UNIVERSITY OF LJUBLJANA
FACULTY OF MATHEMATICS AND PHYSICS
DEPARTMENT OF MATHEMATICS

MATJAŽ KRNC

CENTRALITY MEASURES OF LARGE NETWORKS

DOCTORAL THESIS

ADVISOR:

PROF. DR. RISTE ŠKREKOVSKI

LJUBLJANA, 2015

UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO
ODDELEK ZA MATEMATIKO

MATJAŽ KRNC

MERE SREDIŠČNOSTI VELIKIH OMREŽIJ

DOKTORSKA DISERTACIJA

MENTOR:

PROF. DR. RISTE ŠKREKOVSKI

LJUBLJANA, 2015

Za "atana".

1936 – 2013

So teach us to number our days
that we may get a heart of wisdom.

– Ps 90,12

ABSTRACT

In most networks some edges or vertices are more central than others. To quantify importance of nodes in networks, centrality indices were introduced. For a given structural index, Freeman centralization [52] is a measure of how central a vertex is regarding to how central all the other vertices are with respect to the given index. In the thesis we study several such structural indices like degree, eccentricity, closeness, betweenness centrality, the Wiener index and transmission.

We confirm a conjecture by Everett, Sinclair, and Dankelmann [48] regarding the problem of maximizing closeness centralization in two-mode data, where the number of data of each type is fixed. Intuitively, our result states that among all networks obtainable via two-mode data, the largest closeness is achieved by simply locally maximizing the closeness of a node. Mathematically, our study concerns bipartite networks with fixed size bipartitions, and we show that the extremal configuration is a rooted tree of depth 2, where neighbors of the root have an equal or almost equal number of children.

We determine the maximum value of eccentricity centralization and (some) maximizing networks for the families of bipartite networks with given partition sizes, tree networks with fixed maximum degree and fixed number of vertices, and networks with fixed number of nodes or edges. As a by-product, we introduce and study a new way of enumerating the nodes of a tree.

We also study the centralization of transmission, in particular, we determine the graphs on n vertices which attain the maximum or minimum value. Roughly, the maximizing graphs are comprised of a path which has one end glued to a clique of similar order. The minimizing family of extremal graphs consists of three paths of almost the same length, glued together in one end-vertex.

Group centrality indices, introduced in 1999 by Borgatti and Everett, measure the importance of sets of nodes in networks. We study the notion of group centralization with respect to eccentricity, degree and betweenness centrality measures. For groups of size 2, we determine the maximum achieved value of group eccentricity and group betweenness centralization and describe the corresponding extremal graphs. For group degree centralization we do the same with arbitrary size of group.

For a given integer k, by reduction to maximum domination problem [107], we observe that determining the maximum group degree centralization some k-subset of $V(G)$ is NP-hard. We describe polynomial algorithm with the best-

possible approximation ratio that calculates all centralizations for $1 \leqslant k \leqslant n$ and altogether runs in $O(n^2)$ time. The constructed algorithm is tested on six real-world networks. In results we observe a property of unimodality of group degree centralization for parameter k, which may be a new property for studying networks.

The well studied Wiener index $W(G)$ of a graph G is equal to the sum of distances between all pairs of vertices of G. Denote by $W[\mathcal{G}_n]$ the set of all values of the Wiener index over all connected graphs on n vertices and let the largest interval which is fully contained in $W[\mathcal{G}_n]$ be denoted by W_n^{int}. In the thesis, we show that W_n^{int} is well-defined, it starts at $\binom{n}{2}$, and that both W_n^{int} and $W[\mathcal{G}_n]$ are of cardinality $\frac{1}{6}n^3 + O\left(n^2\right)$ (in other words, most of integers between the smallest value $\binom{n}{2}$ and the largest value $\binom{n+1}{3}$ are contained in W_n^{int} and consequently in $W[\mathcal{G}_n]$).

We describe the above results in separate chapters that are concluded with some further discussion about open problems and future work. In the final chapter we include a short description of our work in progress on betweenness centralization, discuss some of presented results and summarize open problems on extremal graphs.

Math. Subj. Class. (2010): 05C82, 05C35
Keywords: centrality, Freeman centralization, extremal graphs, group centrality

POVZETEK

V večini omrežij so nekatera vozlišča ali povezave pomembnejše od drugih. Pomembnost vozlišč v omrežjih lahko izrazimo z merami centralnosti. Podanemu centralnostnemu indeksu lahko določimo indeks Freemanove centralizacije [52], ki meri relativno centralnost vozlišča v primerjavi s centralnostjno vseh ostalih vozlišč v omrežju. V tej disertaciji analiziramo različne strukturne indekse, kot so stopnja točk, ekscentričnost, centralnost bližine, vmesnostna centralnost, Wienerjev indeks ter totalna razdalja.

Potrdimo domnevo avtorjev Everett, Sinclair in Dankelmann [48] glede maksimiziranja bližinske centralizacije v dvodelnih omrežjih, s podanimi velikostmi biparticij. Trdimo, da je največja vrednost centralizacije bližine (med vsemi dvodelnimi omrežji) dosežena, če lokalno maksimiziramo bližinsko centralnost v neki točki. Izkaže se, da je ekstremalna konfiguracija dosežena v korenskem drevesu globine 2, z dodatnim pogojem, da imajo vsi sosedje od korena skoraj enako stopnjo.

Med drugim določimo maksimizirajočo vrednost ekscentrične centralizacije ter najdemo nekaj maksimizirajočih omrežij za družine dvodelnih grafov s podanimi velikostmi biparticij, dreves fiksne velikosti s podano maksimalno stopnjo, kot tudi splošnih povezanih omrežij pri podanem številu vozlišč ali povezav. Tekom omenjene analize predstavimo tudi nov način enumeracije drevesnih vozlišč.

Totalna razdalja vozlišča v je enaka vsoti vseh razdalj med v ter vsemi drugimi vozlišči v omrežju. Pri analizi centralizacije totalne razdalje določimo grafe na n točkah, ki dosežejo maksimalno ter minimalno vrednost le-tega indeksa. Izkaže se, da so maksimizirajoči grafi sestavljeni iz poti, ki je na enem koncu identificirana s kliko podobne velikosti. Minimizirajoči grafi so sestavljeni iz treh poti podobne velikosti, ki imajo eno krajišče identificirano v skupni točki.

Centralnostni indeksi skupin, vpeljani l. 1999 (Everett in Borgatti [46]), merijo pomembnost izbrane množice vozlišč v omrežju. V disertaciji preučujemo skupinske indekse centralizacije ekscentričnosti, stopnje, ter vmesnostne centralnosti. Za skupine velikosti 2 določimo največje dosežene vrednosti skupinske ekscentričnosti ter skupinske vmesnostne centralnosti, hkrati pa določimo tudi pripadajoče ekstremalne grafe. Podobno določimo tudi za skupinsko centralnost stopnje, neodvisno od velikosti skupine.

Na problem določanja najboljše skupine v smislu skupinske centralizacije stopnje pri podanem omrežju G se osredotočimo tudi algoritmično. Pri podani velikosti skupine k omenjeni problem prevedemo na problem maksimalne k-dominacije [107], ter opazimo da je le-ta $\mathcal{NP}$-težak. Opišemo polinomski algoritem z najboljšim možnim aproksimacijskim koeficientom, ki za vse smiselne velikosti k izračuna centralizacijske vrednosti v skupni časovni zahtevnosti $O(n^2)$. Omenjeni algoritem testiramo na šestih realnih omrežjih. V rezultatih opazimo lastnost unimodalnosti (za parameter k), ki se lahko uporabi kot nova metoda za preučevanje velikih omrežij.

Wienerjev indeks $W(G)$ grafa G je enak vsoti razdalj med vsemi pari vozlišč v G. Z $W[\mathcal{G}_n]$ označimo množico vseh vrednosti Wienerjevega indeksa za družino povezanih omrežij na n vozliščih, pri čemer največji neprekinjen interval iz $W[\mathcal{G}_n]$ označimo z W_n^{int}. V disertaciji pokažemo, da je W_n^{int} smiselno definiran ter se začne v vrednosti $\binom{n}{2}$. Poleg tega pokažemo, da je velikost obeh W_n^{int} ter $W[\mathcal{G}_n]$ vsaj $\frac{1}{6}n^3 + O\left(n^2\right)$, tj. večina vrednosti med $\binom{n}{2}$ ter $\binom{n+1}{3}$ je vsebovana v W_n^{int} (ter posledično tudi v $W[\mathcal{G}_n]$).

Zgornje rezultate predstavimo v ločenih poglavjih, ter jih zaključimo z morebitnimi idejami za prihodnje delo ter odprtimi domnevami. V zaključnem poglavju vključimo kratek povzetek tekočega dela v zvezi z vmesnostno centralizacijo, izpostavimo nekatere predstavljene rezultate ter pregledno povzamemo nekatere odprte domneve na področju ekstremalnih grafov.

Math. Subj. Class. (2010): 05C82, 05C35
Ključne besede: centralnost, Freemanova centralizacija, ekstremalni grafi, skupinska centralnost

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

[83] "Algorithmic approach to degree centralization in large networks" (joint work with M. Karlovčec and R. Škrekovski), *Submitted for publication* (2015).

[84] "Group centralization of network indices" (joint work with R. Škrekovski), *Discrete Appl. Math. (2015), 147–157*.

[85] "Centralization of transmission in networks" (joint work with R. Škrekovski), Discrete Math. (2015), 2412–2420.

[86] "On the Wiener Inverse Interval Problem" (joint work with R. Škrekovski, *To appear in MATCH Commun. Math. Comput. Chem.* (2015).

[87] "Betweenness Centralization Measures for Two-mode Data of Prescribed Sizes" (joint work with J.-S. Sereni, R. Škrekovski, Z. Yilma), *In preparation*.

[88] "Closeness Centralization Measure for Two-mode Data of Prescribed Sizes" (joint work with J.-S. Sereni, R. Škrekovski, Z. Yilma), *Submitted for publication to Netw. Sci. (Camb. Univ. Press, 2015)*.

[89] "Eccentricity of Networks with Structural Constraints" (joint work with J.-S. Sereni, R. Škrekovski Z. Yilma), *Submitted for publication in Ars Math. Contemp.* (2015).

When God created the world,
he did not care about the details.

— R. Škrekovski

ACKNOWLEDGMENTS

I would like to express my very great appreciation to my advisor Dr. Riste Škrekovski, for his valuable and constructive suggestions during the planning and development of my research. Furthermore, he enabled me to work with many great mathematicians, work on interesting problems and make some good friends. I would like to thank my collaborators – M. Karlovčec, J.-S. Sereni, R. Škrekovski and Z. Yilma. Also, comments provided by my PhD committee (J.-S. Sereni, R. Škrekovski and S. Klavžar) were greatly appreciated.

I am grateful for being able to collaborate with friends from Artificial Intelligence Laboratory at Jozef Stefan Institute. The projects and discussions we shared provided me with a sense of real-world networks and related problems.

My special thanks are extended to the friends from my research visits, from which I had the pleasure to learn from, and work with. In addition to the above mentioned, among these are Jernej Azarija, Tomislav Došlić, Zdeněk Dvořák, Tomas Kaiser, František Kardoš, Łukasz Kowalik, Daniel Král', Anita Liebenau, Borut Lužar, Lukáš Mach and Jan Volec.

Naj na koncu izrazim veliko hvaležnost moji ženi Sabini, za neomajno ljubezen in potrpežljivost, moji mami, za požrtvovalno varstvo hčerke Rahele, ter preostali razširjeni družini, za vso podporo.

CONTENTS

1

INTRODUCTION

It is hard to write properly about networks in general, since the term *network* is used in many different contexts. The related field of *network analysis* has grown in various directions, depending on the motivation of the researchers. While the longest tradition of network analysis is credited to the field of social sciences [53], experts that analyze networks come from various other fields, such as mathematics, computer science, anthropology, electrical circuits, project planning, complex systems, transportation systems, communication networks, epidemiology, bioinformatics, hypertext systems, bibliometrics, text analysis, organization theory, genealogical research or event analysis.

As our focus in mathematics is graph theory, we see network analysis as *applied graph theory*. Graph theory is a branch of mathematics that is both topological and combinatorial in nature, and is a comprehensive structural model, or family of models, which is essentially *free from context* and with strong tools already developed for network analysis. Based on this, the notations and terms from network analysis are nowadays mostly unified with terms from graph theory (a big contribution to achieving this is by Hage and Harary [62]).

In most networks some edges or vertices are more central than others. To quantify this intuitive feeling, centrality indices were introduced. First mathematical concept of centrality of graphs was introduced 146 years ago by Jordan [74]. There are many ways to provide a measure of the relative "importance" of a node in a network, thus different motivations lead to different centrality measures that were developed in several fields. Centrality is used in chemistry [73], psychology [2], sociology [72], geography [99], game theory [58], and many other fields. Arguably, the most common branch of centrality functions is based on the distance between the nodes of the network. Standard centrality indices from this branch are betweenness, closeness and eccentricity. Classic applications of these centralities can be found in transportation-network theory [59], communication-network theory [115] and many others. In the last years, even more widespread interest is developing in the field of chemistry [9], electrical circuits [33] and also in the study of food webs [12].

The thesis is structured into two parts. Part one is consisted of some preliminary theory from graph theory and network analysis which are discussed in Chapters 2 and 3, respectively. Among other things, some relevant classes of graphs and structural indices are presented, including some basic historical background that we omit from Chapters 4–9. In the second part we present our main results on several structural indices, such as degree, eccentricity, closeness, betweenness centrality, the Wiener index and transmission.

In Chapter 4 we confirm a conjecture by Everett, Sinclair, and Dankelmann [48] regarding the problem of maximizing closeness centralization in two-mode data, where the number of data of each type is fixed. In particular, our result states that among all networks obtainable via two-mode data, the largest closeness is achieved by simply locally maximizing the closeness of a node. Mathematically, our study concerns bipartite networks with fixed size bipartitions, and we show that the extremal configuration is a rooted tree of depth 2, where neighbors of the root have an equal or almost equal number of children.

In Chapter 5 we determine the maximum value of eccentricity centralization and (some) maximizing networks for the families of bipartite networks with given partition sizes, tree networks with fixed maximum degree and fixed number of vertices, and networks with fixed number of nodes or edges. As a by-product, we introduce and study a new way of enumerating the nodes of a tree, which might be of independent interest.

The transmission of a vertex v in a connected graph G is equal to the sum of distances between v and all other vertices of G. In Chapter 6 we study the centralization of transmission, in particular, we determine the graphs on n vertices which attain the maximum or minimum value. Roughly, the maximizing graphs are comprised of a path which has one end glued to a clique of similar order. The minimizing family of extremal graphs consists of three paths of almost the same length, glued together in one end-vertex.

Group centrality indices, introduced in 1999 by Everett and Borgatti [46], measure the importance of sets of nodes in networks. We study the notion of group centralization with respect to eccentricity, degree and betweenness centrality measures. For groups of size 2, we determine the maximum achieved value of group eccentricity and group betweenness centralization and describe the corresponding extremal graphs. For group degree centralization we do the same with arbitrary size of group.

In Chapter 8, for fixed k, we observe that determining the maximum group degree centralization some k-subset of $V(G)$ is NP-hard. We reduce the problem to maximum domination problem [107], and describe polynomial algorithm with the best-possible approximation ratio that calculates all centralizations for $1 \leqslant k \leqslant n$ and altogether runs in $O(n^2)$ time. The constructed algorithm is tested on six real-world networks. In experiments we observe a property of unimodality of maximum k-group degree centralization.

The well studied Wiener index $W(G)$ of a graph G is equal to the sum of distances between all pairs of vertices of G. Denote by $W[\mathcal{G}_n]$ the set of all values of the Wiener index over all connected graphs on n vertices and let the largest interval which is fully contained in $W[\mathcal{G}_n]$ be denoted by W_n^{int}. In Chapter 9, we show that W_n^{int} is well-defined, it starts at $\binom{n}{2}$, and that both W_n^{int} and $W[\mathcal{G}_n]$ are of cardinality $\frac{1}{6}n^3 + O(n^2)$ (in other words, most of integers between the smallest value $\binom{n}{2}$ and the largest value $\binom{n+1}{3}$ are contained in W_n^{int} and consequently in $W[\mathcal{G}_n]$).

We describe the above results in separate chapters, that are concluded with some further discussion about open problems and future work. Finally, in Chapter 10 we include a short description of some work in progress on betweenness centralization, discuss some of presented results and summarize open problems on extremal graphs.

Part I

PRELIMINARIES

2

GRAPH THEORY

Throughout the thesis, the term *network* is used to refer to the informal concept describing an object composed of elements and interactions or connections between these elements. For example, the Internet is a network composed of vertices (routers, hosts) and connections between these nodes (e.g. fiber cables). The field of graph theory seems to be a natural way to model networks mathematically and some of the same concepts in the intersection of both fields have different names. Hence, some standard graph-theoretic terms (vertex, edge, graph,...) will sometimes be interchanged with those from networks (node, connection, network,...) without any distinction in the meaning.

2.1 BASIC NOTIONS

A *graph* G is an object formed by a set of vertices and a set of edges that connect pairs of vertices. The vertex set and edge set of a graph G are denoted by $V(G)$ and $E(G)$, respectively, while their cardinalities are usually denoted by n or m, respectively. The two vertices joined by an edge are called its *endvertices*. If two vertices are joined by an edge, they are adjacent and we call them *neighbors*. Graphs can be undirected or directed. In undirected graphs, the order of the endvertices of an edge is irrelevant. An undirected edge joining vertices $u, v \in V(G)$ is denoted by $\{u, v\}$. Whenever we are dealing with directed graphs, each directed edge has a destination and an origin. An edge with destination $v \in V(G)$ and origin $u \in V(G)$ is represented by an ordered pair (u, v). For convenience, an edge $\{u, v\}$ or (u, v) will usually be denoted by uv. In an undirected graph, uv and vu both stand for $\{u, v\}$, while in a directed graph uv is short for (u, v). Graphs that can have directed edges as well as undirected edges are called mixed graphs, but such graphs are encountered rarely and we will not discuss them in the thesis. An edge is called *a loop*, if both its endpoints are the same. The graph is *simple*, if it is undirected and does not contain loops or multiple edges. Unless stated otherwise, we will assume that our graphs are simple.

The *degree* of a vertex v, denoted by $\deg(v)$, represents the number of its neighbours. The set of neighbors of v is denoted by $N(v)$. If the graph G is directed the *in-degree* of $v \in V(G)$, denoted by $\deg^-(v)$, corresponds to the number of edges with destination v. The *out-degree* of $v \in V(G)$, denoted by $\deg^+(v)$, is the number of edges in $E(G)$ with origin in v. An undirected graph is called *regular* if all of its vertices have the same degree, and *k-regular* if that degree is equal to k. Vertices of degree one are called *leaves*. The maximum, minimum and average degree of an undirected graph G are denoted by $\Delta(G)$, $\delta(G)$ and $\overline{d}(G)$, respectively.

A graph G' is a *subgraph* of the graph G if $V(G') \subseteq V(G)$ and $E(G') \subseteq E(G)$. It is an *induced subgraph* if $E(G')$ contains all edges $e \in E(G)$ that join vertices in $V(G')$. The induced subgraph of G with vertex set $X \subseteq V(G)$ is denoted by $G[X]$. If F is a subset of $E(G)$, then $G - F$ denotes the graph obtained from G by deleting all edges in F. If C is a proper subset of $V(G)$, then $G - C$ denotes the graph obtained from G by deleting all vertices in C and their incident edges.

A *path* in a graph G is a sequence of edges which connect a sequence of pairwise distinct vertices. The *distance* $d_G(u,v)$ in G of two vertices u, v is the length of the shortest path with endpoints u and v. If no such path exists, we set $d(u,v) = \infty$. The distance $d(v,U)$ between a vertex v and a set of vertices $U \subseteq V(G)$ is defined as $d(v,U) = \min_{u \in U} d(v,u)$. The *diameter* of G, denoted $\operatorname{diam}(G)$, stands for the greatest distance between any two vertices in G, i.e. $\max_{u,v \in V(G)} d(u,v)$. Two graphs G and H are *isomorphic*, if there exists a bijection $f : V(G) \to V(H)$, such that all edges are preserved, i.e. $uv \in E(G) \Leftrightarrow f(u)f(v) \in E(H)$. The above relation is denoted with $G \simeq H$.

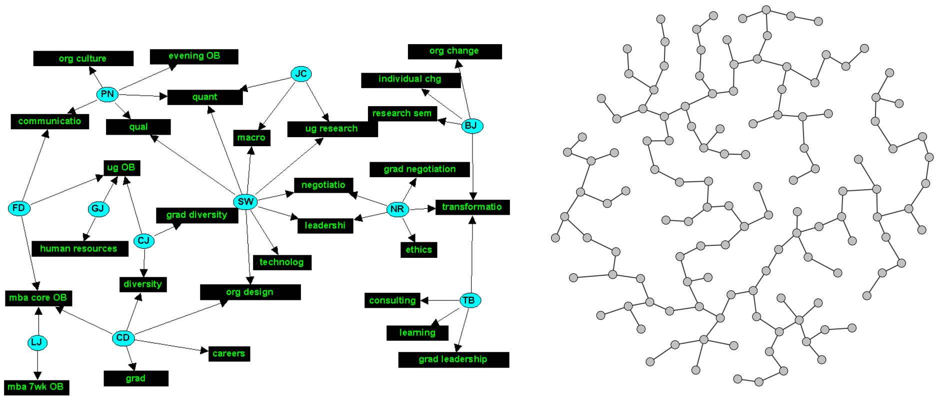
Let $\gamma(G)$ be a cardinality of a minimum set that dominates graph G (also known as *domination number*), i.e. it is the smallest integer, such that

$$\exists S \in \binom{V(G)}{\gamma(G)} : \bigcup_{v \in S} (N(v) \cup \{v\}) = V(G).$$

A function f is said to be *unimodal* if locally there is only a single highest value in f. If the graph under consideration is not clear from the context, these and other notations will sometimes be augmented by specifying the graph as an index. For example, $\deg_H(v)$ denotes the degree of vertex v in the graph H.

2.2 RELEVANT CLASSES OF GRAPHS

The family of all connected graphs $\mathcal{G}_n$ represents the most basic class of graphs – the collection of all possible connected graphs with vertex-set of order n. We now define some classes of graphs that we use throughout the thesis.



(a) a bipartite network (b) a tree

Figure 2.1: Examples of a directed bipartite graph (a) and a tree (b).

Bipartite graphs

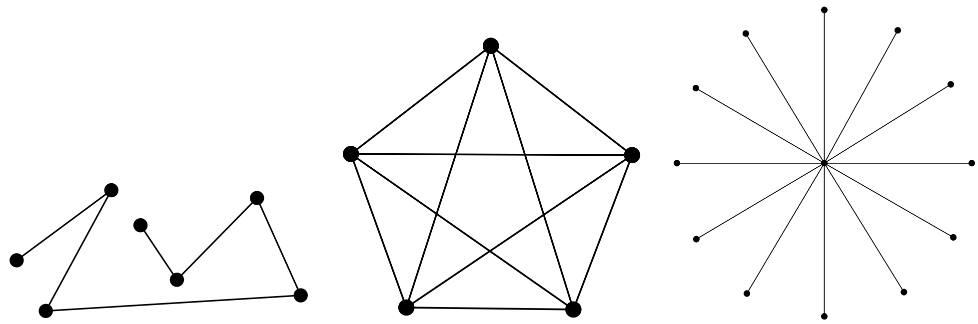
If we can partition the vertex set of an (undirected) graph into two parts, such that no edge connects two vertices from the same part, we call such graph a *bipartite* graph. Denote by $\mathscr{B}(n_0, n_1)$ the family of all bipartite networks on fixed bipartition sizes n_0 and n_1. Bipartite networks are important also in social sciences [21], where the word *two-mode data* is usually used instead. An example of a directed bipartite graph can be observed on Fig. 2.1a, where circles represent faculties, squares are courses, and arrows indicate which faculty chose which courses.

Complete graphs and complete bipartite graphs

Another standard class of graphs are the complete graphs. The *complete* graph K_n is a graph on n vertices, where any pair of distinct vertices is connected by an edge. The complete bipartite graph $K_{a,b}$ is a graph on $a + b$ vertices, where any pair of vertices from distinct bipartitions is connected by an edge. Among $\mathcal{G}_n$, graph K_n maximizes the number of edges and some other centrality indices, described in Section 3.3 on page 21. An example of a complete graph can be observed on Fig. 2.2b.

Trees and forests

An undirected graph is *a tree*, if for any pair of its distinct vertices, there exists the unique path between them. Note that all trees are bipartite. An example



(a) a path (b) a complete graph (c) a star

Figure 2.2: An example of P_7 (a), K_5 (b) and S_{12} (c).

can be observed on Figure 2.1b. An undirected graph is a *forest*, if for any pair of its distinct vertices, there exists at most one path between them. As no cycles are allowed in trees, they are members of bipartite graphs by definition. The family of trees, usually denoted with $\mathcal{T}$ is very important in graph theory and most of graph families described below are members of the trees.

Paths and stars

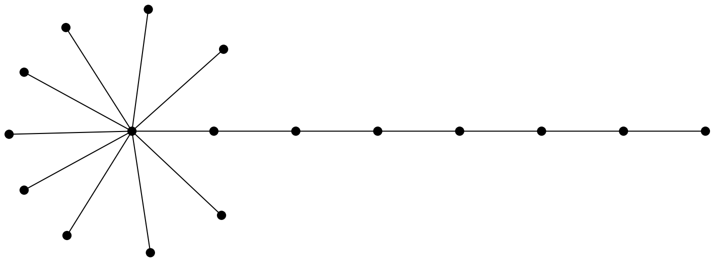
A very simple subfamily of trees are *paths*. Graph on n vertices is a path P_n, if it contains two leaves, while $n-2$ of its remaining vertices are of degree 2. Among $\mathcal{G}_n$, graph P_n maximizes diameter, radius, Wiener index, and some other graph indices. An example of a path is shown on Fig. 2.2a.

A *star* S_n is a tree on $n+1$ vertices, consisted of a vertex connected to n leaves, i.e. $S_n \simeq K_{1,n}$. Among graphs from $\mathcal{G}_n$, star is a tree with maximum number of leaves. As we observe later, among graphs from $\mathcal{G}_n$, the star graphs are the maximizing family for several topological indices. An example of a star S_{12} is depicted in Figure 2.2c.

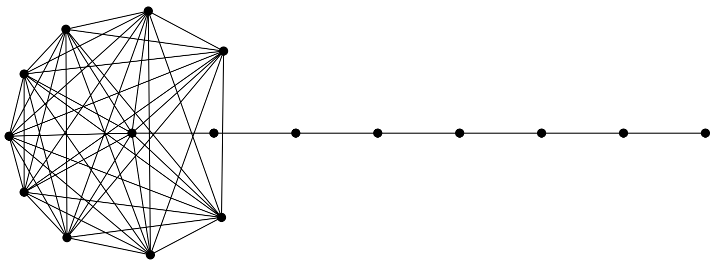
The graph families that follow are a mixture of path, star and complete graphs.

Dandelion graphs and comets

Let $D(n,l)$ be the *Dandelion* graph on n vertices, consisted of the star S_{n-l} and a path P_l, on vertices $p_0, p_1, \ldots, p_{l-1}$, where p_0 is identified with a star center. The notion of Dandelion graphs will be important in Chapters 7 and 9. An example of $D(17,8)$ is shown in Fig. 2.3a.



(a) Graph $D(17, 8)$.



(b) Graph $C(17, 8)$.

Figure 2.3: Examples of a Dandelion graph and a comet.

The family of *comets* looks similar as the family of dandelion graphs. Let $C(n, l)$ be the *comet* on n vertices, consisted of a complete graph K_{n-l+1} and a path P_l, on vertices $p_0, p_1, \ldots, p_{l-1}$, where p_0 is identified with a vertex from K_{n-l+1}. An example of $C(17, 8)$ is shown in Fig. 2.3b.

The bipartite family of graphs H

Later on in the thesis we will also consider rooted trees on fixed partition sizes n_0 and n_1 that are of depth two and have nicely distributed degrees of all non-root vertices. Formally, they are described as follows (see Everett et al. [48]).

Definition 2.1. Let $H(v; n_0, n_1)$ be the connected bipartite tree with node bipartition (A_0, A_1) such that

- $|A_i| = n_i$ for $i \in \{0, 1\}$;
- there exists a node $v \in A_0$ such that $N_G(v) = A_1$; and
- $\deg(w) \in \left\{1 + \left\lceil \frac{n_0 - 1}{n_1} \right\rceil, 1 + \left\lfloor \frac{n_0 - 1}{n_1} \right\rfloor \right\}$ for all nodes $w \in A_1$.

The node v is called the *root* of $H(v; n_0, n_1)$. The family of graphs $H(\cdot, \cdot, \cdot)$ will play an important role in Chapter 4. An example of a graph $H(0, 18, 14)$ can be observed on Fig. 2.4.

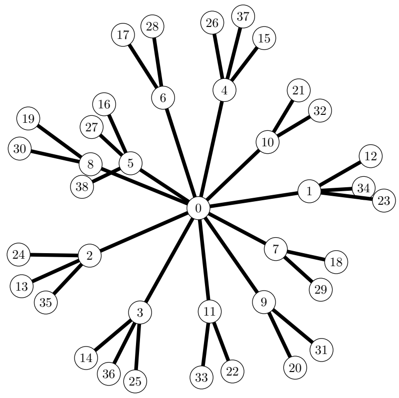


Figure 2.4: A graph $H(0, 18, 14)$.

3

NETWORK THEORY

For many decades, social networks have been a fundamental subject of study in social sciences. With the rapid growth of the internet and the world wide web in recent years, many large-scale on-line based social networks appeared, and many large-scale social network data became easily available, thereby providing an important source of materials for analysis [43, 109, 110, 118].

As the connections among the nodes can exhibit complicated patterns, a network can be a complex structure. Still, most networks appearing in nature, despite their diversity, follow some universal organizing principles. When studying complex networks, one challenge is to develop simplified measures that capture some elements of the structure in an understandable way. Let us describe some of those here.

3.1 BASIC TERMS AND PROPERTIES

As mentioned in the introduction, a network is often conveniently modeled by a graph: nodes (vertices) represent individual objects and connections (edges) represent the relationships between pairs of these objects. In the thesis we will freely interchange terms vertex/node and graph/network, without any meaningful difference. We work on simple unweighted networks: our network only tells us, for a given (binary) relation R, which pairs of individual are in relation according to R.

Degree distribution

One property that ignores any patterns among different nodes and just look at each node separately is the *degree distribution*. If one zooms in onto a node and ignores all other nodes, the only thing one can see is how many connections the node has, i.e., the degree of the node. The degree distribution of a graph G is a probability distribution that maps integers $[1, \Delta_G]$ to the real interval

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×	....	×	×	....	....	....	....	....
2. Miss Laura Mandeville	×	×	×	....	×	×	×	×	....	....	....	....	....	....
3. Miss Theresa Anderson	....	×	×	×	×	×	×	×	×	....	....	....	....	....
4. Miss Brenda Rogers	×	....	×	×	×	×	×	×	....	....	....	....	....	....
5. Miss Charlotte McDowd	....	....	×	×	×	....	×	....	....	....	....	....	....	....
6. Miss Frances Anderson	....	....	×	....	×	×	....	×	....	....	....	....	....	....
7. Miss Eleanor Nye	....	....	....	....	×	×	×	×	....	....	....	....	....	....
8. Miss Pearl Oglethorpe	....	....	....	....	....	×	....	×	×	....	....	....	....	....
9. Miss Ruth DeSand	....	....	....	....	×	....	×	×	×	....	....	....	....	....
10. Miss Verne Sanderson	....	....	....	....	....	....	×	×	×	....	....	×	....	....
11. Miss Myra Liddell	....	....	....	....	....	....	....	×	×	×	....	×	....	....
12. Miss Katherine Rogers	....	....	....	....	....	....	....	×	×	×	....	×	×	×
13. Mrs. Sylvia Avondale	....	....	....	....	....	....	×	×	×	×	....	×	×	×
14. Mrs. Nora Fayette	....	....	....	....	....	×	×	....	×	×	×	×	×	×
15. Mrs. Helen Lloyd	....	....	....	....	....	....	×	×	....	×	×	×	....	....
16. Mrs. Dorothy Murchison	....	....	....	....	....	....	....	×	×	....	....	....	....	....
17. Mrs. Olivia Carleton	....	....	....	....	....	....	....	....	×	....	×	....	....	....
18. Mrs. Flora Price	....	....	....	....	....	....	....	....	×	....	×	....	....	....

DEEP SOUTH

source: Davis et al. [38]

Figure 3.1: A non-trivial part of an adjacency matrix of two-mode network. Rows represent the participation of women in social events reported in Old City Herold.

$[0, 1]$. It contains information about the probability of the degree of a randomly chosen node in a network. In other words, we can define it as

$$P_{\text{deg}}(k) = \text{ a fraction of nodes of degree } k \text{ in } G.$$

An example of the degree distributions for two-mode networks from Figures 2.1a and 3.1 above can be observed on Figure 3.2.

Power-law degree distributions and scale-free Networks

If the number of nodes $P_{\text{deg}}(k)$ of degree k is given by $P_{\text{deg}}(k) \propto k^{-\gamma}$ for some $\gamma > 1$, we call such degree distribution *a power-law* distribution.

When we increase the scale or units by which we measure a distribution of a power-law network by a constant factor, the shape of the distribution remains unchanged except for the multiplicative constant. For this reason, networks with a power-law distribution are also called *scale-free* networks.

The fact that power law degree distribution is very common among large real-world networks shows that real networks are not "random" (see a related study by Faloutsos et al. [49]). From a power-law degree distribution, it follows that such scale-free networks contain many vertices with a degree that greatly exceeds the average. Those are often called "hubs", and are thought to serve specific purposes in their networks, although this depends greatly on the domain. Examples of scale-free networks include citation graphs [124], phone call graphs [1], the Web [76, 27, 11, 69, 90, 100] (also see Figure 3.3), on-line

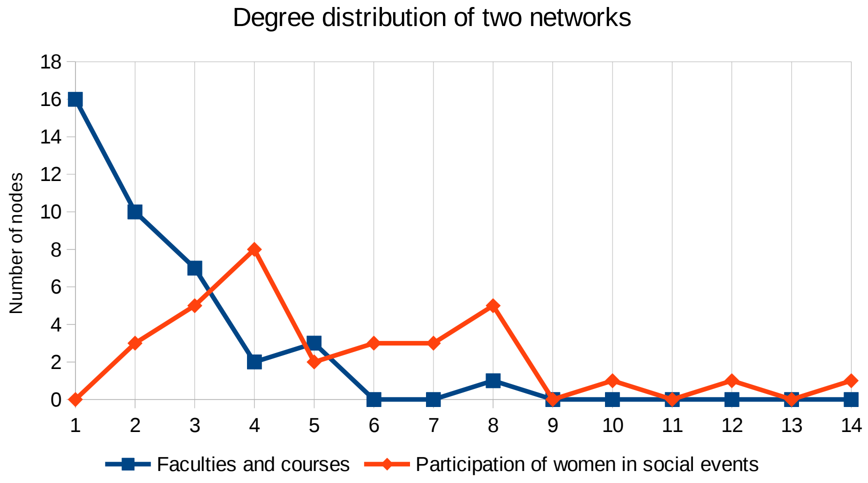


Figure 3.2: The non-normalized degree distributions of graphs from Figures 2.1a and 3.1.

Network	The degree exponent γ
The web graph [11]	$\gamma_{in} = 2.1$
	$\gamma_{out} = 2.4$
Autonomous systems [49]	$\gamma = 2.4$

Table 3.1: Real-world examples of different values of the degree exponent in two networks.

social networks [30], the Internet [49], click-stram data [18] and many others. For most networks, the degree exponent falls inside an interval $\gamma \in [2, 3]$. Two real-world examples can be observed in Table 3.1.

Network Diameter

As noted above, a network diameter is the maximal distance between any pair of nodes in a network. By this definition, disconnected graphs have infinite diameter. In 2001, Tauro et al. [132] introduced a notion of an *effective diameter*, which is the smallest distance at which at least 90% of all *connected* pairs of nodes can be reached.

Small-world Networks

Many networks have an interesting mathematical property so that most nodes can be reached from every other by a small number of steps, and at the same time most nodes are not neighbors of one another. Those graphs are also called *small-world networks* (see [139]). Formally, a small-world network is defined to be a network with

$$d \propto \log n,$$

where d is an expected distance between two randomly chosen nodes and n is the number of nodes in a graph. Since any isolated vertex force a diameter of a whole network to be infinite, big real-world networks often use more robust measure of effective diameter.

In the context of a social network, constant or small expected distance between two randomly chosen nodes implies the so-called *small world phenomenon*. In 1998, Watts and Strogatz [139] published the first network model on the small-world phenomenon, which we further discuss in Section 3.2. The characteristics of small-world network can be observed in social networks, the connectivity of the Internet, Web and many other networks [4, 103, 5, 19, 27, 32, 139].

On Figure 3.3 one can observe a representation of the most known small-world networks – World Wide Web by B. Lyon [100], where different colors represent different continents of origin.

Clustering coefficient

In most real-world networks, in particular social networks, nodes tend to create highly connected groups characterized by a relatively high edge density; the edge-probability in such a group is greater than the average probability of a random edge between two nodes (see [66, 139]). The *clustering coefficient* is a measure correlated to a number of triangles in the network, i.e., sets of fully connected triples of nodes. Let G be our network and let Δ_o and Δ_c be the number of induced P_3 and K_3 in G, respectively. The clustering coefficient of G is defined as:

$$C(G) = \frac{\Delta_c}{\Delta_c + \Delta_o}.$$

Similarly, *clustering coefficient of a vertex v* is defined as

$$C(v) = \frac{\Delta'(v)}{\binom{\deg(v)}{2}},$$

where $\Delta'(v)$ represents a number of triangles that contains a vertex v. A further generalizations to weighted, directed and bipartite networks are presented in [116, 39, 98, 3].

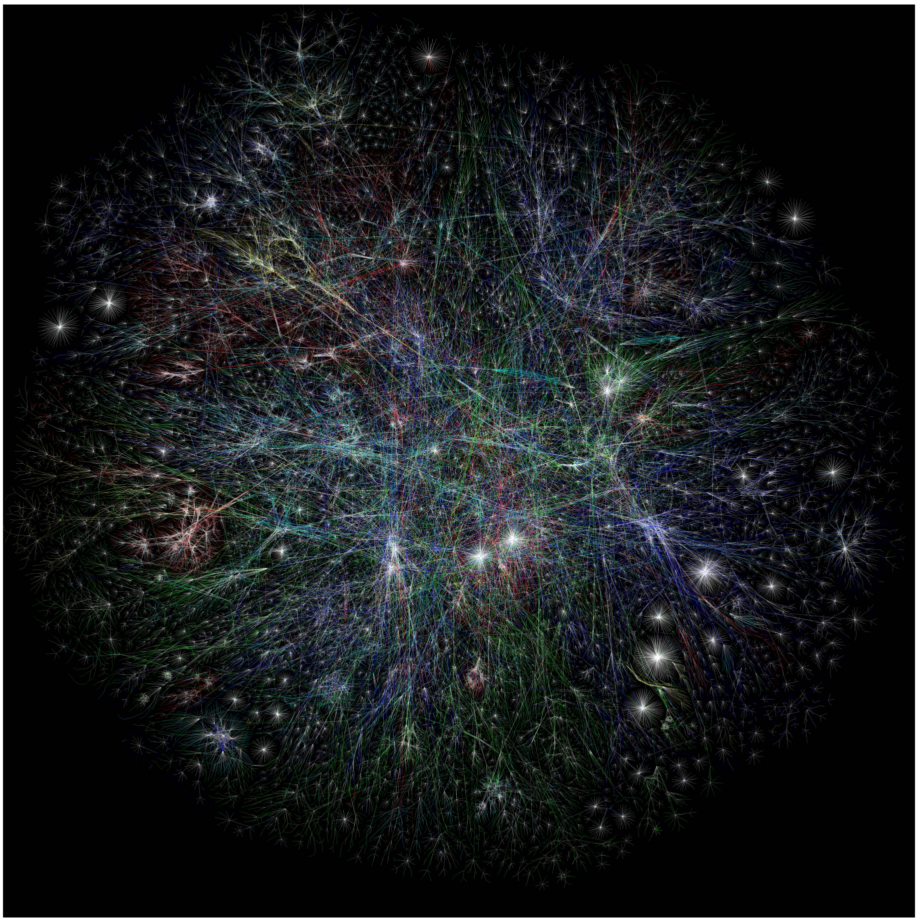


Figure 3.3: The web-graph [100]; one of most known small-world networks. Different colors represent different continents of origin.

Studies show that clustering coefficient is significantly higher in real networks than in random networks of the same degree distribution (the model that generates such random graphs is described in [105]). In real networks, it has further been observed [41, 122] that for a vertex v clustering coefficient $C(v)$ decreases as $\deg(v)$ increases. Moreover, $C(v)$ scales as a power law, i.e. $C(v) \propto \deg(v)^{-1}$. Consider a social network in which nodes are people and links are acquaintance relationships between people. One interpretation of this phenomenon is that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs. An example of clustering in social networks are people that tend to form small groups in which everyone knows almost everyone else (communities). These are often organized or hierarchically nested.

One can observe community-like sets of nodes tend in various networks. For example, they correspond to functional modules in biological networks [123], organizational units in social networks [111], and scientific disciplines in collaboration networks between scientists [57].

Graph motifs and graphlets

Network motifs [104, 6] are basic building blocks of complex networks:¸

> *Network motifs are sub-graphs that repeat themselves in a specific network or even among various networks. Each of these sub-graphs, defined by a particular pattern of interactions between vertices, may reflect a framework in which particular functions are achieved efficiently. Indeed, motifs are of notable importance largely because they may reflect functional properties.* (from [143])

The idea is to enumerate and count occurrences of all possible subgraphs of a given graph G up to a small number of nodes. The subgraphs considered are usually of size up to $n = 5$ nodes, as the computation for larger values of n is unfeasible. The frequencies of motifs are stored in a vector and compared to those of a random graph with the same degree distribution (the model that generates such random graphs is described in [105]). Finally, one can extract motifs that occur significantly more frequently in real-world network than in the random-generated graph.

A variant of network motifs are so-called *graphlets*, introduced by Pržulj et al. [120], that requires all small subgraphs to be induced. On Figure 3.4 one can observe all 30 non-isomorphic graphs on $n \in \{2, 3, 4, 5\}$ nodes (denoted by $G_0, G_1, \ldots, G_{29}$) that correspond to all 72 non-isomorphic rooted graphlets.

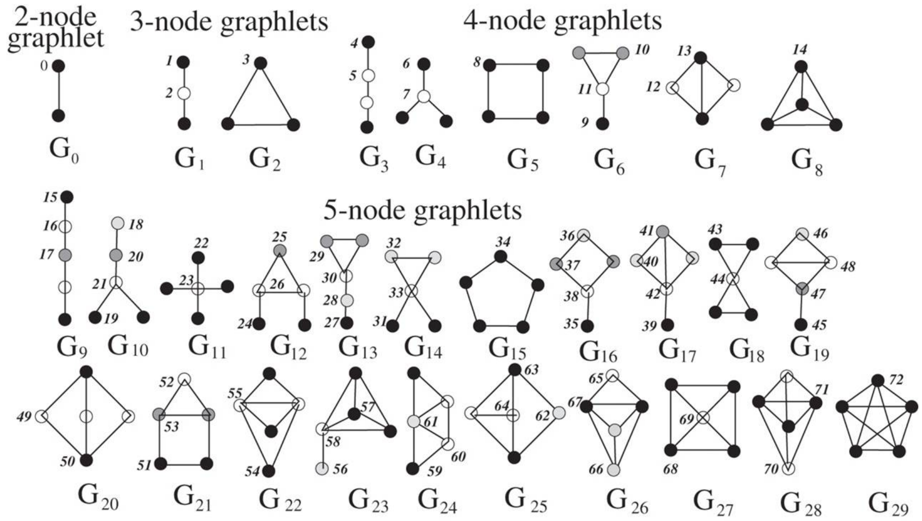


Figure 3.4: All 30 non-isomorphic graphs on $n \in \{2, 3, 4, 5\}$ nodes (denoted by $G_0, G_1, \ldots, G_{29}$) that correspond to all 72 non-isomorphic rooted graphlets.

3.2 MODELS

Another interesting field in network analysis is generation of networks with given specific properties. Given the set of properties or limitations that we want the resulting network to have, one needs to construct a procedure that will construct the desired graph. In the base part of our thesis, in particular in Chapters 4–8, we will deal with a similar task; trying to find the structure of graphs that maximizes some very special centrality-related property. In this section we present some of the most common models for generating real-world like networks and describe the appropriate global properties that can be observed in each of them.

Erdős-Rébyi random graph

In graph theory, the two most known models for generating random graphs (namely $\mathcal{G}_{n,p}$ and $\mathcal{G}_{n,m}$) are named after Paul Erdős and Alfréd Rényi. They first introduced the $\mathcal{G}_{n,p}$ model in 1959 (see [44]), while the other model was independently introduced by Edgar Gilbert [56]. In the model denoted $\mathcal{G}_{n,p}$, an edge between each pair from n nodes is placed independently of the other edges, with a fixed probability p. The other variant (denoted $\mathcal{G}_{n,m}$) is to fix number of nodes and edges (n and m respectively) and then uniformly at random choose a graph from a set of all possible graphs on n nodes and m edges. In the probabilistic method, these models can be used to provide a rigorous

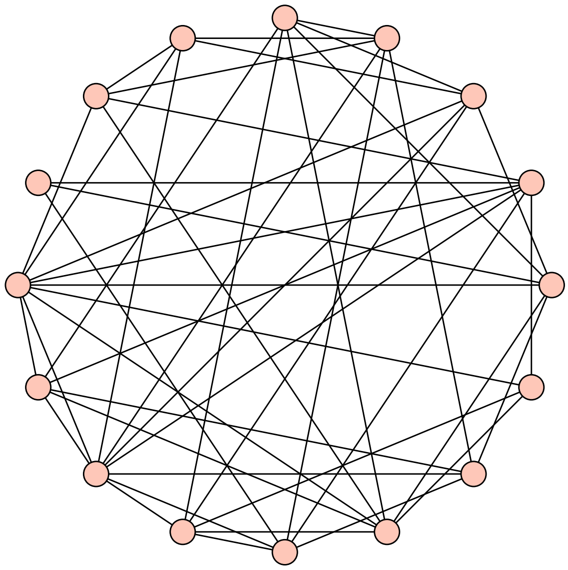


Figure 3.5: Erdős-Rébyi random graph with $n = 16$ and $m = 48$.

definition of what it means for a property to hold for almost all graphs, or to prove the existence of graphs satisfying various properties. A close correspondence exists between both models and in practice most theorems hold for both variants. An example of Erdős-Rébyi random graph can be observed on Fig. 3.5.

One can show that the average shortest path length of a random graph grows with the number of nodes n as $O(\log \log n)$ and that the diameter increases asymptoticly as $O(\log n)$ [31]. Furthermore, the degree distribution of Erdős-Rébyi random graph follows a binomial distribution with mean $\overline{d}$ [4] (note that $\overline{d}$ stands for an average degree of a network). When assessing the clustering coefficient of some network, graph motifs vector or graphlet degree-vector, one usually compare it to the values from random graph with the same degree distribution. The model that generates such random graphs is described in [105].

Although Erdős-Rébyi random graph model is the fundamental one for graph generation, the resulting graphs do not have some important properties observed in many real-world networks. Since they have a constant and independent probability of two nodes being connected, their clustering coefficient is artificially small. Furthermore they do not account for the formation of hubs. Finally, the degree distribution of a general Erdős-Rébyi graphs converges to a Poisson distribution, rather than a power law observed in many real-world, scale-free networks. To address some of the above mentioned limitations, the following models were designed.

Watts and Strogatz model

The Watts–Strogatz model is a random graph generation model that produces graphs with small-world properties, including short average path lengths and high clustering. As its name suggests, it was proposed Watts and Strogatz [139].

For a positive integer n and an even integer k, let $C(n,k)$ denote the graph obtained from a cycle C_n by adding the additional edges uv, whenever $d_{C_n}(u,v) \leqslant k$. The Watts–Strogatz graph $WS(n,k,p)$ can be constructed from $C(n,k)$ by replacing each edge in $C(n,k)$ with probability p by a randomly chosen edge. An example of Watts and Strogatz random graph can be observed on Fig. 3.6a.

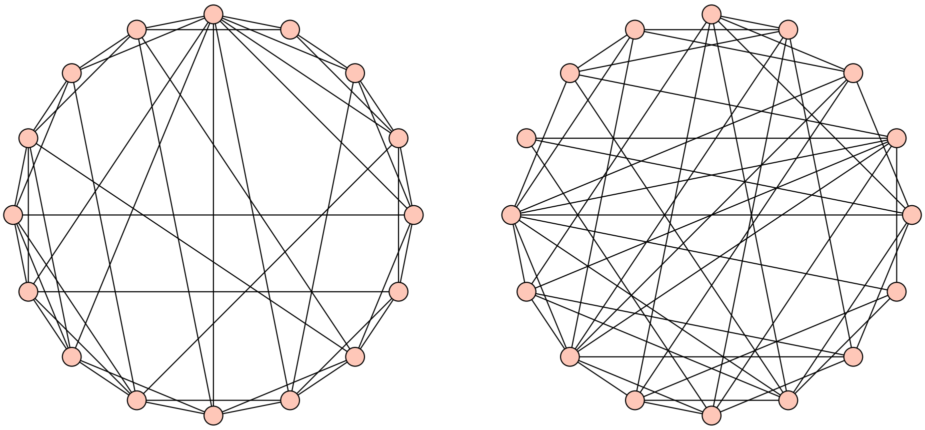
Barabási–Albert model

Unfortunately, Watts and Strogatz model is also unable to generate networks that have power-law degree distribution. By using preferential attachment method, the *Barabási–Albert model* [4] is an algorithm that indeed generates random scale-free networks. The algorithm constructs a network by adding new nodes to the network one at a time.

The algorithm begins with some initial connected graph of m_0 nodes. Each new node is connected to $m \leqslant m_0$ existing nodes with a probability that is proportional to the number of links that the existing nodes already have. This results in a desired phenomenon so that vertices of high degree tend to quickly accumulate even more neighbors, while those with only a few links are unlikely to be chosen as an endpoint of a new edge. An example of Barabási–Albert random graph can be observed on Fig. 3.6b.

3.3 NETWORK CENTRALITY

In most networks some edges or vertices are more central than others. To quantify this intuitive feeling, centrality indices were introduced. First mathematical concept of centrality of graphs was introduced 146 years ago by Jordan [74]. There are many ways to provide a measure of the relative "importance" of a node in a network, thus different motivations lead to different centrality measures that were developed in several fields. Centrality is used in chemistry [73], psychology [2], sociology [72], geography [99], game theory [58], transportation-network theory [59], communication-network theory [115] and



(a) Watts and Strogatz random graph with $n = 16$ and $k = 5$ and $p = \frac{1}{2}$.

(b) Barabási–Albert random graph with $n = 16$ and $m = 4$.

Figure 3.6: Graphs on 16 nodes and 48 edges, generated by different random models.

many other fields. Arguably, the most basic centrality measure is the degree of the node, and the most common branch of other centrality functions is based on the distance between the nodes of the network. Standard centrality indices from this branch are betweenness, closeness and eccentricity. For detailed definitions and discussion on various centrality indices, we refer the reader to the work of Brandes and Erlebach [25], Bavelas [13, 15], Koschützki et al. [82], Proctor and Loomis [119], Seeley [127]. In the last years, even more widespread interest is developing in the field of chemistry [9], electrical circuits [33] and also in the study of food webs [12].

In this section we outline some fundamental concepts in centrality theory. First consider one of fundamental properties that all centrality measures share.

Definition 3.1. Let G be an (undirected) graph. A function $f : V(G) \to \mathbb{R}$ is called a *structural index* if the following condition is satisfied:

$$\forall v \in V(G) : G \simeq H \Longrightarrow f_G(v) = f_H(\phi(v)),$$

where ϕ is an isomorphism between G and H.

Clearly, two isomorphic networks should always attain the same centrality values on their nodes, therefore any centrality index is required to be a structural index. Thus, relation $\leqslant$ with respect to any centrality measure induces a semi-order on the vertex-set of a graph. For a given centrality index, it is natural to ask if it satisfies the *rule of monotonicity*, defined as follows.

Definition 3.2. Let G be an (undirected) graph. A centrality measure $f: V(G) \to \mathbb{R}$ on G *satisfies the rule of monotonicity* if u is "more central" than v whenever $f(u) > f(v)$, for any two nodes u and v of G.

In next two subsections, we present two important aspects of centrality measures, namely group centrality and Freeman centralization.

3.3.1 *Group Centrality*

In 1999, Everett and Borgatti [46] introduced the concept of *group centrality* which enables researchers to answer questions such as "how central is the engineering department in the informal influence network of this company?" or "among middle managers in a given organization, which are more central, the men or the women?" With these measures we can also solve the inverse problem: given the network of ties among organization members, how can we form a team that is maximally central? In [46], the authors introduced group centrality for measures of degree, closeness and betweenness centrality, which we use in this paper. In 2006, Borgatti introduced another important group centrality measure (usually called KPP) that is motivated by *key players problem* (see [20]). In his paper he focused on finding a set of vertices for the purpose of optimally diffusing something through the network by using selected vertices as seeds, or for maximally fragmenting the network by removing the key nodes. Interestingly, Borgatti claims that previously mentioned group closeness and betweenness are not proper tools to define KPP centrality. He therefore used tools like graph fragmentation and information entropy to define KPP centrality. To make the distinction between a vertex or a group centrality index precise, the following definition is used in the thesis.

Definition 3.3 (Group Centrality Measure)**.** Let G be a (directed) graph. The class of *group centrality indices* is the class of functions $f : 2^{V(G)} \to \mathbb{R}$ that are invariant under isomorphism.

Several more concepts of centrality with respect to some subset of vertices have been introduced throughout last decade. Those can easily be mistaken for a group centrality measure, but are in fact an extended type of vertex-centrality indices. In 2003, Smith and White [141] introduced a measure called *personalization* that shows, how central an individual is according to given subset R (group of important people) in given social network. In 2005, *subgraph centrality* has been introduced by Estrada and Rodríguez-Velázquez [45], and characterizes the participation of each node in all subgraphs in a network, which is calculated from the spectra of the adjacency matrix of the network. In the same year, Everett and Borgatti in [47], introduced another measure (i.e. *core centrality*), where they evaluate the extent to which a network revolves around a core group of nodes. Finally, very recently Bell [17] introduced the concept called *subgroup centrality*, where centrality (of one vertex) is calculated only on restricted set of vertices. Let us remark that all four mentioned centralities in principle measure importance of an individual vertex (with respect to some conditions) and are different from group centrality, proposed in [46]. In Section 3.4 we present some relevant group centrality indices that we use in the thesis.

It should be noted that the centrality ratio or difference of two vertices cannot generally be used to quantify how much more central one vertex is comparing to the other one. For example, having 350 connections on Facebook was an average degree for US in 2014. On the other hand, having the same degree in coauthorship network would be quite impressive. To be able to compare such vertices from different context, a proper normalization of a given centrality index is needed. Arguably, the most standard way to do this is by using *Freeman centralization* that we define in the following subsection.

3.3.2 *Network Centralization*

In his study, Freeman [52] realized that despite all defined vertex-centrality indices, there was a need for *graph centrality* measure based on differences in point centrality. He defined a *centralization* index that can be used in combination with any vertex-centrality to determine to what extent some vertex in network stands out from others in terms of given centrality index. Furthermore he used this approach to compare different graphs, depending on their highest centralization scores. The general definition of centralization for graphs assigns a centralization measure F_1 to any existing centrality measure F.

Definition 3.4. If F is a vertex-centrality measure, then set

$$\overline{F_1}(G,v) = \sum_{u \in V(G)} \left(F(v) - F(u)\right). \tag{3.1}$$

In order to compare centralization values of graphs with different sizes, Freeman used a normalized formula, dividing the expression (3.1) by the theoretically largest such sum of differences in any graph from the given class of graphs. In general, for centrality index F, its centralization measure is thus defined as

$$F_1(G,v) = \frac{\overline{F_1}\,(G,v)}{\max_{G' \in \mathcal{G}_n} \max_{v' \in V(G')} \overline{F_1}(G',v')}. \tag{3.2}$$

There is some interest in studying the extremal graphs for centralization of various centrality indices. Freeman argued that the centralizations for degree centrality, betweenness centrality and closeness centrality attain their maximum if and only if G is the star network. The statement was later proved in

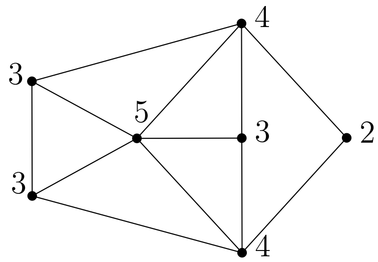


Figure 3.7: An example of a graph and its degrees.

detail by Everett et al. [48], therefore in (3.2) one can replace G' by a star S_{n-1}. In 2006, Butts [28] studied the extremal values of degree centralization among all graphs on n vertices. Further results on extremal graphs for centralization can be found Chapters 4–8.

Whenever graph G is known from the context, we omit it from the notions of centrality or centralization.

3.4 RELEVANT STRUCTURAL INDICES

We now give an overview of structural indices that we use in the thesis. For each of them, we also define their Freeman centralization versions, as well as their group centrality version.

3.4.1 *Degree Centrality*

Historically the first and most basic centrality measure is the *degree centrality* of a vertex v that is simply defined as the degree $\deg(v)$ of a vertex v if the considered graph is undirected (Figure 3.7). In a directed case there are two possibilities of defining the degree centrality: the in-degree centrality $\deg^-(v)$ and the out-degree centrality $\deg^+(v)$. The degree centrality can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information). The degree centrality for directed graphs is also applicable whenever the graph represents something like a voting result. These networks represent a static situation and we are interested in the vertex that has the most direct votes or that can reach most other vertices directly. The degree centrality is a local measure, because the centrality value of a vertex is only determined by the number of its neighbors. Applying Freeman centralization to degree centrality yields the following definition from [52].

Definition 3.5. Let G be a graph on n vertices and m edges and let D be the degree centrality function. The *degree centralization* is defined as:

$$\overline{D_1}(G, v) = \frac{n \cdot \deg_G(v) - 2m}{(n-1)\,(n-2)}.$$

Here we used the fact that among all graphs in $\mathcal{G}_n$, the graph that maximizes $\overline{D_1}$ is the star S_n, and that $\max_{v \in V(G')} \overline{D_1}(S_n, v) = (n-1)\,(n-2)$, where v is the center of the star. In 2006, Butts [28] studied bounds for an unnormalized version of degree centralization in graphs with different densities. He showed that roughly half of the region of conceivable degree centralization scores is actually feasible, and that the geometry of the feasible region alters with the graph size.

We now give the definition of *group centrality* measure by Everett and Borgatti [46], which basically counts the cardinality of the neighborhood of a given set of vertices.

Definition 3.6. Let G be a graph and let $S \subseteq V(G)$. *Group degree centrality* is defined as

$$\mathrm{GD}(S, G) = \left| \bigcup_{v \in S} N(v) \setminus S \right|.$$

3.4.2 *Eccentricity*

The aim of the next centrality index is to determine a node that minimizes the maximum distance to any other node in the graph.

A first paradigm for location based problems on the minimization of transportation costs was introduced by Weber [140] in 1909. However, no significant progress was made before 1960, when facility location emerged as a research field. Facility location analysis deals with the question of finding optimal locations for one or more facilities in a given environment. Applications include transportation-networks [64] and communication-networks [26]. The spatial location of facilities often takes place in the context of a given transportation, communication, or transmission system, which may be represented as a network for analytic purposes. As an example, consider the problem of determining the location for an emergency facility such as a hospital. The main objective of such an emergency facility location problem is to find a site that minimizes the maximum response time between the facility and the site of a possible emergency.

The motivation described above corresponds to a centrality measure known as *eccentricity* (or *group eccentricity*). The eccentricity $e(v)$ of a node v in a connected network G is the maximum distance (in the network) between v and

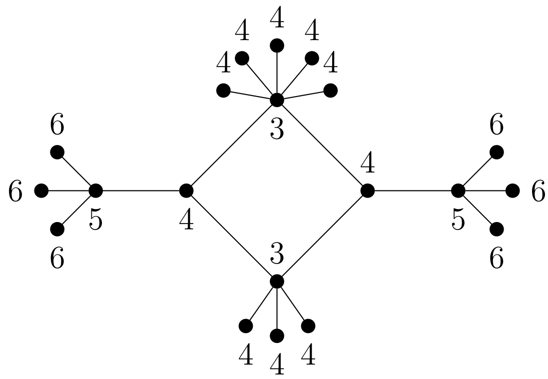


Figure 3.8: A network with the eccentricity of each node.

u, over all nodes u of G. For a disconnected network, all nodes are defined to have infinite eccentricity. To state this formally:

$$e(v) = \begin{cases} \max\left\{d_G(v,u) \, : \, (u,v) \in V(G)^2\right\} & ; G \text{ is connected,} \\ \infty & ; \text{otherwise.} \end{cases}$$

The *center* (or *Jordan center* [138]) of a network is the set of all nodes of minimum eccentricity, i.e. $\{v \in V(G); \forall u \in V(G)\, e(v) \leqslant e(u)\}$[101]. The set of points from the center (also called *central* or *median points*) is denoted with $m(G)$. An example of graph with its eccentricity values is shown on Figure 3.8.

Based on eccentricity, Hage and Harary [63] proposed a corresponding centrality measure.

Definition 3.7. Let G be a graph and let $v \in V(G)$. The centrality measure of eccentricity is defined as

$$E(v) = \frac{1}{e_G(v)}.$$

The reciprocal of the eccentricity value is convenient, since it obeys the rule of monotonicity from Definition 3.2. Let us state the Freeman centralization of eccentricity [52].

Definition 3.8. A Freeman centralization of eccentricity measure is the *eccentricity centralization*, given by

$$\forall v \in V(G), \quad \overline{E_1}(G,v) = \frac{\sum_{u \in V(G)}(E_G(v) - E_G(u))}{n-1}.$$

Here we use the fact that the star graph S_n achieves the maximal value of E_1, and that $\max_{v \in V(S_n)} E_1(S_n, v) = n - 1$.

In order to determine the location of more sites that minimizes the maximum response time between closest facility and the site of a possible emergency, group eccentricity [84] measure need to be used.

Definition 3.9. Let G be a graph and let $C \subseteq V(G)$. The *group eccentricity* of a set C is defined as

$$\mathrm{GE}(G, C) = \frac{1}{\max_{x \in V(G)} d(x, C)},$$

where $d(x, C) = \min_{c \in C} d(x, c)$.

3.4.3 *Betweenness Centrality*

A popular indicator of a node's centrality is the *betweenness centrality*, introduced by Anthonisse [7] and popularized by Freeman [51]. As the name suggests, betweenness measures the extent to which a given vertex is situated in paths between pairs of vertices. This measure is of particular importance in communication networks where a vertex v can attain a certain level of importance, responsibility, or status by controlling the flow of information across the network. In the simplest case, where there is a unique path between any pair of vertices (i.e., when the network is a forest), a vertex v can block or facilitate information flow between all pairs of vertices $\{x, y\}$ for which the unique path between x and y passes through v. It would be natural, then, to define $B(v) = B(v; G)$, the betweenness centrality index of the vertex v in the graph G, to be the number of exactly such pairs $\{x, y\}$. Given a graph G and a pair of vertices $s, t \in V(G)$, let $\sigma_{s,t}$ be the number of (s, t)-paths of length $d(s, t)$. For a vertex $v \notin \{s, t\}$, let $\sigma_{s,t}(v)$ be the number of shortest (s, t)-paths that go through v. So, the expression $\sigma_{s,t}(v)/\sigma_{s,t}$ gives the proportion of shortest (s, t)-paths that pass through v. Let us now formally define betweenness centrality.

Definition 3.10. The *betweenness centrality* of a vertex v is then defined as follows:

$$B(v) = B(v; G) = \sum_{s \neq v \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}.$$

In this regard, $B(v)$ measures how much information flows through v. However, it is usually also interesting to know the relative importance of the vertex v. For example, if a graph G is highly symmetric, it may be the case that a vertex controls the most flow without exerting dominance; that is, all vertices have a roughly equal share of responsibility and no one vertex can be considered as the hub. Therefore, it becomes important to define the *betweenness centralization* index, B_1, of a vertex as a measure of its dominance over other vertices in information flow.

Definition 3.11. Let G be a graph and let $v \in V(G)$. The *betweenness centralization* is defined as

$$\overline{B}_1(G, v) = \frac{\sum_{u \in v} B(v) - B(u)}{\frac{1}{2}(n-1)^2(n-2)}.$$

For a highly symmetric graph G it is clear that $B_1(v, G)$ would be close to zero, no matter the choice of v. On the other hand, it is an easy calculation to show that the central vertex of the star $K_{1,n}$ has betweenness centralization index $n\binom{n}{2}$ which, as shown in [52], is the maximum betweenness centralization value that can be attained by vertices belonging to graphs on $n+1$ vertices (one can also appeal here to the very idea of being a *hub*). The fact that the star maximizes B_1 in the family of $\mathcal{G}_n$ is also the reason for the denominator in definition above. The central vertex of $K_{1,m}$ is also easily seen to maximize $\overline{B}_1(\cdot, \cdot)$ if one fixes the number of edges, m.

Definition 3.12. Let G be a graph and let $C \subseteq V(G)$. Let $\sigma_{u,v}(C)$ be the number of geodesics connecting u to v passing through some vertex of C. Then, the *group betweenness centrality* of C is given by

$$\mathrm{GB}(C) = \sum_{\{u,v\} \subseteq V(G) \setminus C} \frac{\sigma_{u,v}(C)}{\sigma_{u,v}}.$$

3.4.4 *Closeness Centrality*

Suppose that we want to place one or more service facilities, e.g., a shopping malls, such that the total distance from nearest shopping mall to all inhabitants in the region is minimal. This would make the chosen locations as convenient as possible for most inhabitants. In network analysis the centrality index based on this concept is called *closeness centrality*. Closeness centrality measures how close a node is to all other nodes in the network: the smaller the total distance from a node v to all other nodes, the more important the node v is. It was introduced by Bavelas [15]. On Figure 3.9 the total sum of distances from each vertex to all others is displayed.

Definition 3.13. Let G be a graph and let v be one of its vertices. *Closeness centrality* is defined as

$$C(v) = \frac{1}{\sum_{u \in V(G)} d(u, v)}.$$

Since the star graph S_n again attains maximal value of $\overline{C_1}$, it is easy to calculate that $\max_{v \in V(S_n)} \overline{C_1}(S_n, v) = \frac{1}{2} - \frac{1}{4n-6} = \frac{n-2}{2n-3}$. Hence we can state the Freeman centralization of closeness.

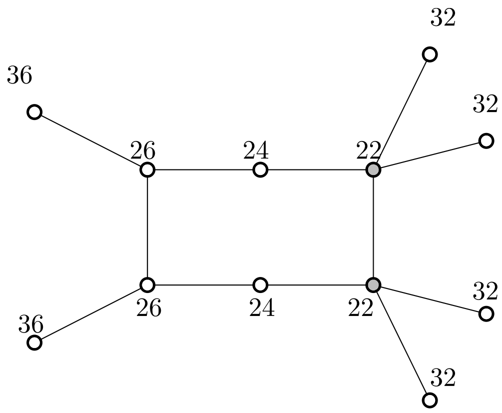


Figure 3.9: The total sum of distances from each vertex to all others.

Definition 3.14. Let G be a graph and let $v \in V(G)$. Then *closeness centralization* is defined as

$$\overline{C_1}(G,v) \quad = \tfrac{2n-3}{n-2} \cdot \sum_{u \in V(G)} C(v) - C(u).$$

The case of positioning multiple shopping malls in the motivation above corresponds to the notion of *group closeness centrality*, introduced by Everett and Borgatti [46].

Definition 3.15. Let G be a graph and let $C \subseteq V(G)$. Then, the *group closeness* centrality is defined as

$$\mathrm{GC}(C) = \frac{1}{\sum_{v \in V(G)} d_G(v, C)}.$$

In [46], authors also mention some other variants of group closeness centrality by giving three different definitions of $d_G(v, U)$. In the thesis we only use version with the usual meaning of $d(v, U)$, i.e. $d(v, U) = \min_{u \in U} d(v, u)$.

3.4.5 *Wiener index and Transmission*

The Wiener index $W(G)$, introduced by Wiener [142], is a graph index defined for connected graph G as the sum of the lengths of shortest paths between all unordered pairs of vertices in G, formally

$$W(G) = \sum_{u,v \in V(G)} d_G(u, v).$$

It is the oldest topological index related to molecular branching and based on its success, many other topological indices correlated to distance matrix of chemical graphs have been developed subsequently to Wiener's work. Wiener index was at first used for predicting the boiling points of paraffins [142], but later a strong correlation between Wiener index and other chemical or physical properties of a compound was found, such as critical points in general [131], the density, surface tension, and viscosity of compounds liquid phase [125] and the van der Waals surface area of the molecule [60].

There are some recent papers on Wiener index of trees [67], common neighborhood graphs [34, 78] and line graphs [144, 77, 80, 79]. Finding graph extremals for Wiener index and its derivatives is nicely summarized in a recent survey by Gutman et al. [145]. It is easy to conclude that among connected graphs on n vertices, minimal and maximal values of Wiener index are $\binom{n}{2}$ and $\binom{n+1}{3}$ obtained at K_n and P_n, respectively. In the class of trees, both extremal graphs are S_n and P_n with Wiener values $(n-1)^2$ and $\binom{n+1}{3}$, respectively. These and many other bounds for the Wiener index are presented in [145, 81].

It is easy to conclude that among connected graphs on n vertices, minimal and maximal values of Wiener index are $\binom{n}{2}$ and $\binom{n+1}{3}$ obtained at K_n and P_n, respectively. In the class of trees, both extremal graphs are S_n and P_n with Wiener values $(n-1)^2$ and $\binom{n+1}{3}$, respectively.

Transmission of a particular vertex $v \in V(G)$ (in some literature also called *farness* or *vertex-Wiener index*) is defined as a sum of the lengths of all shortest paths between chosen vertex and all other vertices in G, i.e.

$$W(v) = \sum_{u \in V(G) \setminus \{v\}} d_G(u, v).$$

In order to define a Freeman centralization to Wiener index of a vertex, notice that

$$W_1(v) = \sum_{u \in V(G) \setminus \{v\}} (W(v) - W(u)) = n \cdot W(v) - 2W(G).$$

Using the results from Chapter 6 we can also determine the maximizing graph for W_1 and calculate the denominator of Wiener (or transmission) centralization.

Definition 3.16. Let G be a graph and let $v \in V(G)$. *Transmission centralization* of a vertex v is defined by

$$\overline{W_1}(v) = \frac{n \cdot W(v) - 2W(G)}{f(n)},$$

where

$$f(n) = \begin{cases} \frac{5}{24}n^3 - \frac{3}{4}n^2 + \frac{19}{24}n - \frac{1}{4} & ; n \text{ is odd}, \\ \frac{5}{24}n^3 - \frac{3}{4}n^2 + \frac{2}{3}n & ; n \text{ is even}. \end{cases}$$

3.4.6 *Other Centrality Measures*

Let us also mention two other centrality measures that we do not discuss in the thesis.

Eigenvector centrality assigns relative scores to all vertices in the graph based on the concept that connections to vertices with high centrality contribute more to the score of the vertex in question than connections to low-scoring vertices. For a vertex v, the eigenvector centrality $\mathrm{Eig}(v)$ is defined as

$$\mathrm{Eig}(v) = \frac{1}{\lambda} \sum_{u \in N(v)} \mathrm{Eig}(u),$$

where λ is a constant. For different values of λ this system can have many different solutions. However, since all the eigenvector values must be positive, not all solutions are admissible. In particular, by the Perron–Frobenius theorem [117], only the greatest eigenvalue results in the desired centrality measure. A variant of the eigenvector centrality measure is Google's PageRank. Another closely related centrality measure is the Katz centrality, that we describe now.

Katz centrality was introduced by Leo Katz in 1953 [75] and is a generalization of the transmission. It is a distance-type centrality measure that measures influence of a node v by taking into account the total number of walks between v and all other vertices, where the long walks have less contribution than the short ones, depending on attenuation factor α, formally

$$C_{\mathrm{Katz}}(v) = \sum_{k=1}^{\infty} \sum_{u \in V(G)} \alpha^k \left(A^k\right)_{u,v}$$

where $(A)_{u,v}$ is the value of adjacency matrix of G that corresponds respectively to the row and the column of vertices u and v.

Part II

SOME RESULTS ON CENTRALITY

4

CLOSENESS CENTRALIZATION FOR TWO-MODE NETWORKS OF PRESCRIBED SIZES

In this chapter we focus on closeness centrality, which measures how close a node is to all other nodes in the network. It is one of standard centrality that assigns more importance to nodes with smaller total distance to all other nodes. We confirm a conjecture by Everett, Sinclair, and Dankelmann [48] regarding the problem of maximizing closeness centralization in two-mode data, where the number of data of each type is fixed. Intuitively, our result states that among all networks obtainable via two-mode data, the largest closeness is achieved by simply locally maximizing the closeness of a node. Mathematically, our study concerns bipartite networks with fixed size bipartitions, and we show that the extremal configuration is a rooted tree of depth 2, where neighbors of the root have an equal or almost equal number of children.

4.1 BASIC NOTIONS

We will work on simple unweighted networks: our network only tells us, for a given (binary) relation $E(G)$, which pairs of nodes are in relation according to $E(G)$.

Formally, for a node v of a network G, the *closeness* of v is defined to be

$$C_G(v) := \frac{1}{\sum_{u \in V(G)} d_G(v,u)}, \tag{4.1}$$

where $d_G(u,v)$ is the length of a shortest path in G between nodes u and v. We shall use the shorthand $W_G(v) := \sum_{u \in V(G)} d_G(v,u)$. In both notations, we may drop the subscript when there is no risk of confusion. Various closeness-based measures have been developed and/or discussed by Bavelas [14], Beauchamp [16], Botafogo et al. [22], Nieminen [114], Moxley and Moxley [108], Sabidussi [126], Valente and Foreman [135], Nieminen [114].

While centrality measures compare the importance of a node within a network, the associated notion of *centralization*, as introduced by Freeman [52], allows us to compare the relative importance of nodes within their respective networks (see 3.3.2 on page 24). The closeness centralization of a node v in a network G is given by

$$C_1(v, G) := \sum_{u \in V(G)} \big[C(v) - C(u)\big]. \tag{4.2}$$

Further, we set $C_1(G) := \max\{C_1(v; G) : v \in V(G)\}$.

It is important to note that the parameter C_1 is really tailored to compare the centralization of nodes in different networks. If only one network is involved, then one readily sees that maximizing $C_1(v; G)$ over the nodes of a network G amounts to minimizing W_G. Indeed, suppose that G is a network and v a node of G such that $W_G(v) \leqslant W_G(u)$ for every $u \in V(G)$. Then, for every node x of G it holds that

$$\begin{aligned} C_1(v;G) - C_1(x;G) &= (n-1)\left(\frac{1}{W_G(v)} - \frac{1}{W_G(x)}\right) - \left(\frac{1}{W_G(x)} - \frac{1}{W_G(v)}\right) \\ &= n\left(\frac{1}{W_G(v)} - \frac{1}{W_G(x)}\right) \\ &\geqslant 0. \end{aligned}$$

In what follows, we use the following notation. The *star network* of order n, sometimes simply known as an *n-star* [134], is the tree on $n+1$ nodes with one node having degree n. The star network is thus a complete bipartite network with one part of size 1. Over all networks with a fixed number of nodes, the closeness is maximized by the star network.

Theorem 4.1 (48)**.** *If G is a network with n nodes, then*

$$C_1(u; S_{n-1}) \geqslant C_1(G),$$

where u is the node of S_{n-1} of maximum degree.

Everett, Sinclair, and Dankelmann [48] considered the problem of maximizing centralization measures for two-mode data. In this context, the relation studied links two different types of data (e.g., persons and events) and we are interested in the centralization of one type of data only (e.g., the most central person). Thus the network obtained is *bipartite*: its nodes can be partitioned into two parts so that all the edges join nodes belonging to different parts. An example of a real-world two-mode network N on 89 edges with partition sizes $|P_1| = 18$ and $|P_2| = 14$, borrowed from Davis et al. [38] is depicted on Figure 4.1. On the figure, one can observe a frequency of interparticipation of a group of women in social events in Old City, 1936. On Tables 4.1 and 4.2, one can observe closeness centralization for partitions P_1 and P_2 and notice that closeness centrality (and hence centralization) is maximized at "Mrs. Evelyn Jefferson" and the event from "September 16th", respectively.

$v \in P_1$	$C_N(v)$	$C_1(v, N)$
Mrs. Evelyn Jefferson	0.01667	0.07779
Miss Theresa Anderson	0.01667	0.07779
Mrs. Nora Fayette	0.01667	0.07779
Mrs. Sylvia Avondale	0.01613	0.06058
Miss Laura Mandeville	0.01515	0.02930
Miss Brenda Rogers	0.01515	0.02930
Miss Katherine Rogers	0.01515	0.02930
Mrs. Helen Lloyd	0.01515	0.02930
Miss Ruth DeSand	0.01471	0.01504
Miss Verne Sanderson	0.01471	0.01504
Miss Myra Liddell	0.01429	0.00160
Miss Frances Anderson	0.01389	−0.01110
Miss Eleanor Nye	0.01389	−0.01110
Miss Pearl Oglethorpe	0.01389	−0.01110
Mrs. Dorothy Murchison	0.01351	−0.02311
Miss Charlotte McDowd	0.01250	−0.05555
Mrs. Olivia Carleton	0.01220	−0.06530
Mrs. Flora Price	0.01220	−0.06530

Table 4.1: Nodes from the group of women and their closeness values.

$v \in P_2$	LABEL ON FIG. 4.1	$C_N(v)$	$C_1(v, N)$
September 16th	P8	0.01923	0.15984
April 8th	P9	0.01786	0.11588
March 15th	P7	0.01667	0.07779
May 19th	P6	0.01562	0.04445
February 25th	P5	0.01351	−0.02311
April 12th	P3	0.01282	−0.04529
April 7th	P12	0.01282	−0.04529
June 10th	P10	0.01250	−0.05555
September 26th	P4	0.01220	−0.06530
February 23rd	P11	0.01220	−0.06530
June 27th	P1	0.01190	−0.07459
March 2nd	P2	0.01190	−0.07459
November 21st	P13	0.01190	−0.07459
August 3rd	P14	0.01190	−0.07459

Table 4.2: Nodes from the partition of social events events from 1936, reported in *Old City Herald*, and their closeness values.

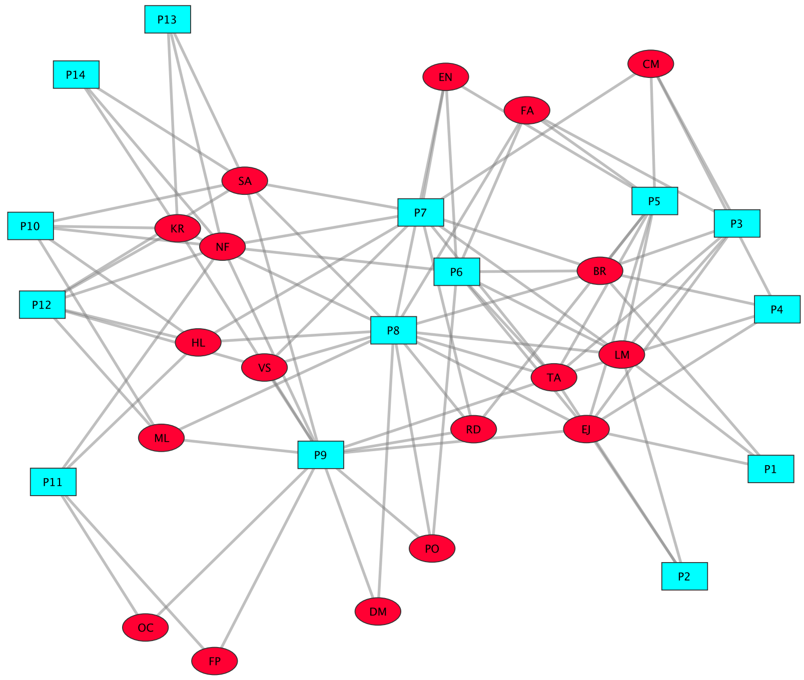


Figure 4.1: A two-mode network N on 89 edges with partition sizes $n_0 = 18$ and $n_1 = 14$. The network represents the participation of a given set of people in the social events from 1936 reported in the Old City Herald, where rectangles represent social events while circles represent women, labeled by their initials (see Tables 4.1 and 4.2).

Everett et al. formulated an interesting conjecture, which was later proved by Sinclair [128]. To state it, we first need a definition.

Definition 4.2. Let $H(v; n_0, n_1)$ be the connected bipartite tree with node bipartition (A_0, A_1) such that

- $|A_i| = n_i$ for $i \in \{0, 1\}$;
- there exists a node $v \in A_0$ such that $N_G(v) = A_1$; and
- $\deg(w) \in \left\{1 + \left\lceil \frac{n_0-1}{n_1} \right\rceil, 1 + \left\lfloor \frac{n_0-1}{n_1} \right\rfloor \right\}$ for all nodes $w \in A_1$.

The node v is called *the root* of $H(v; n_0, n_1)$.

The aforementioned conjecture was that the pair $(H(v; n_0, n_1), v)$ is an *extremal pair* for the problem of maximizing *betweenness centralization* in bipartite networks with a fixed sized bipartition into parts of sizes n_0 and n_1. Recall that for two-mode data, we are only interested in one type of data: in graph-theoretic terms, we look only at nodes that belong to the part of size n_0, and we want to know which of these nodes has the largest closeness in the network. In other words, letting A_0 be the part of size n_0 of $V(G)$, we want to determine $\max\{C_1(u;G) : u \in A_0\}$.

A similar study for the centrality measure of eccentricity is described in Chapter 5. Everett et al. [48] also suggested that the same pair is extremal for closeness and eigenvector centralization measures. In this chapter, we confirm the conjecture for the closeness centralization measure. That is, we prove that the pair $H(v;n_0,n_1)$ is extremal for the problem of maximizing closeness centralization in bipartite networks with parts of size n_0 and n_1, where v is the root.

4.2 BIPARTITE NETWORKS WITH FIXED NUMBER OF NODES

In this section, we present the proof of the main theorem from this chapter, which confirms the claim by Everett et al. [48] about graph $H(\cdot;n_0,n_1)$ being extremal for closeness centralization measure on the class of bipartite networks with fixed bipartition sizes.

Theorem 4.3. *Let G be a bipartite network with node parts A_0 and A_1 of sizes n_0 and n_1, respectively. Then for each $v \in A_0$,*

$$C_1(u;H(u;n_0,n_1)) \geqslant C_1(v;G).$$

To prove Theorem 4.3, suppose that G is a bipartite network with bipartition (A_0,A_1) where $|A_i| = n_i$ for $i \in \{0,1\}$, and u is a node in A_0 such that $C_1(u;G) \geqslant C_1(v;H(v;n_0,n_1))$. We prove that this inequality must actually be an equality by showing that any such extremal pair $C_1(u;G)$ must satisfy the following three properties:

(P1) G is a tree;

(P2) $\deg_G(u) = n_1$; and

(P3) $|\deg_G(w_1) - \deg_G(w_2)| \leqslant 1$ whenever $w_1, w_2 \in A_1$.

Property (P1) is relatively straightforward to check and so is (P3) if we assume that (P2) holds. Thus the majority of the discussion below will be devoted to proving that (P2) holds, which we do last. For convenience, we define V to be $V(G)$.

We start by establishing (P1); namely, that the network G is a tree. Assume, for the sake of contradiction, that G is not a tree and let T be a breadth-first-search tree of G rooted at u. Note that $W_G(u) = W_T(u)$ and $W_T(x) \geqslant W_G(x)$ for any node $x \in V(G)$. In addition, there exist at least two nodes for which the above inequality is strict. It follows that $C_1(u;T) > C_1(u;G)$, a contradiction.

We now establish that (P3) holds if (P2) does. Thus we know that G is a tree and we assume that $N_G(u) = A_1$, therefore also all nodes from $A_0 \setminus \{u\}$ are leaves. Suppose, for the sake of contradiction, that there exist nodes $w_1, w_2 \in A_1$ such that $\deg(w_1) \geqslant \deg(w_2) + 2$. Let z be a neighbor of w_1 different from u and consider the network G' obtained by deleting the edge w_1z and replacing it with w_2z. Note that $W_{G'}(u) = W_G(u)$ and that $W_{G'}(x) = W_G(x)$ unless $x \in N_G[w_1] \cup N_G[w_2]$, that is unless x belongs to the closed neighborhood of either w_1 or w_2. So

$$C_1(u; G') - C_1(u; G) = \sum_{x \in N_G[w_1] \cup N_G[w_2]} \frac{1}{W_G(x)} - \sum_{x \in N_G[w_1] \cup N_G[w_2]} \frac{1}{W_{G'}(x)}. \quad (4.3)$$

Now, let $\{u, z, x_1, \dots, x_t\} = N_G(w_1)$ and $\{u, y_1, \dots, y_s\} = N_G(w_2)$ where, by assumption, $t > s$.

Recalling that G is a tree, observe that the following hold for every $i \in \{1, \dots, t\}$ and every $j \in \{1, \dots, s\}$ (for better illustration, see Figure 4.2).

(i). $W_{G'}(x_i) = W_G(x_i) + 2$;

(ii). $W_{G'}(y_j) = W_G(y_j) - 2$;

(iii). $W_G(y_j) = W_G(x_i) + 2(t - s + 1) > W_G(x_i) + 2$;

(iv). $W_{G'}(z) = W_G(z) + 2(t - s) > W_{G'}(z)$;

(v). $W_{G'}(w_1) = W_G(w_1) + 2$; and

(vi). $W_{G'}(w_2) = W_G(w_2) - 2$.

From (i)–(iii), we infer that for any $j \in \{1, \dots, s\}$,

$$\frac{1}{W_{G'}(x_j)} + \frac{1}{W_{G'}(y_j)} < \frac{1}{W_G(x_j)} + \frac{1}{W_G(y_j)},$$

and similarly by (v) and (vi),

$$\frac{1}{W_{G'}(w_1)} + \frac{1}{W_{G'}(w_2)} < \frac{1}{W_G(w_1)} + \frac{1}{W_G(w_2)}.$$

Thus the right side of (4.3) is greater than

$$\frac{1}{W_G(z)} - \frac{1}{W_{G'}(z)} + \sum_{j=s+1}^{t} \frac{1}{W_G(x_j)} - \frac{1}{W_{G'}(x_j)},$$

which is positive by (i) and (iv). This contradiction shows that (P3) holds provided (P2) does.

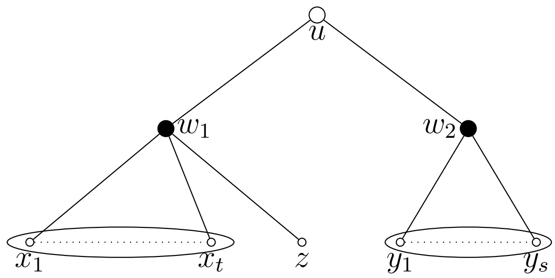


Figure 4.2: The subtree of G induced by $N_G[w_1] \cup N_G[w_2]$.

To complete the proof, it remains to prove that (P2) holds. First, if $n_1 = 1$, then the tree G must be an n_0-star, hence the second property is satisfied. Now consider the case where $n_1 = 2$. Then there is precisely one node x that is adjacent to both nodes in A_1. Moreover, $W_G(x) \leqslant W_G(w)$ if $w \in A_0$ since, if $w \in A_0 \setminus \{x\}$ then $W_G(w) \geqslant 2(n_0 - 1) + 4 = 2n_0 + 2$ while $W_G(x) = 2 + 2(n_0 - 1) = 2n_0 + 1$. Thus $u = x$ and hence $\deg_G(u) = n_1 = 2$, as wanted.

From now on, we assume that $n_1 \geqslant 3$. As in the proof of (P3), we argue that if (P2) does not hold then $C_1(u; G)$ can be increased by altering the network G. In this case, however, we find it necessary to use our assumption that $C_1(u; G)$ itself is at least as large as $C_1(v; H(v; n_0, n_1))$. This shall allow us to have a lower bound on $C_1(u; G)$, thanks to the next lemma.

Lemma 4.4. $C_1(v; H(v; n_0, n_1)) \geqslant \frac{n_1 - 1}{2(2n_1 - 1)}$.

Proof. We establish the inequality via a direct computation. Unfortunately, the expressions involved force a lengthy computation.

We set $m := n_0 - 1$ and we write $m = pn_1 + r$ where $0 \leqslant r < n_1$. Let us now calculate $W(x)$ for each node x of $H(v; n_0, n_1)$.

1. $W(v) = n_1 + 2m$.
2. Consider the neighbors of v. There are
   a) r neighbors x for which $W(x) = \lceil m/n_1 \rceil + 1 + 2(n_1 - 1) + 3(m - \lceil m/n_1 \rceil)$; and
   b) $n_1 - r$ neighbors x for which $W(x) = \lfloor m/n_1 \rfloor + 1 + 2(n_1 - 1) + 3(m - \lfloor m/n_1 \rfloor)$.
3. Consider the nodes at distance two from v. There are
   a) $r \lceil m/n_1 \rceil$ nodes x for which $W(x) = 1 + 2\lceil m/n_1 \rceil + 3(n_1 - 1) + 4(m - \lceil m/n_1 \rceil)$; and
   b) $(n_1 - r) \lfloor m/n_1 \rfloor$ nodes x for which $W(x) = 1 + 2\lfloor m/n_1 \rfloor + 3(n_1 - 1) + 4(m - \lfloor m/n_1 \rfloor)$.

Since $\lfloor m/n_1 \rfloor = (m-r)/n_1$ and, for $r > 0$, we have $\lceil m/n_1 \rceil = (m + n_1 - r)/n_1$. It follows that

$$C_1(v) = \frac{n_1 + m}{n_1 + 2m} - \frac{rn_1}{3mn_1 - 2m + 2n_1^2 - 3n_1 + 2r} - \frac{n_1(n_1 - r)}{3mn_1 - 2m + 2n_1^2 - n_1 + 2r} - \frac{r(m + n_1 - r)}{4mn_1 - 2m + 3n_1^2 - 4n_1 + 2r} - \frac{(n_1 - r)(m - r)}{4mn_1 - 2m + 3n_1^2 - 2n_1 + 2r} \tag{4.4}$$

$$\geqslant \frac{n_1 + m}{n_1 + 2m} - \frac{n_1^2}{3mn_1 - 2m + 2n_1^2 - 3n_1 + 2r} - \frac{n_1 m}{4mn_1 - 2m + 3n_1^2 - 4n_1 + 2r}, \tag{4.5}$$

where we used that $n_1 > 0$ to derive (4.5).

As is seen from (4.4), if n_1 is fixed and n_0 tends to infinity (hence, so does m), then $C_1(v)$ approaches $1/2 - n_1/(4n_1 - 2) = \frac{n_1 - 1}{4n_1 - 2}$.

Let us now subtract $\frac{n_1 - 1}{4n_1 - 2}$ from the right side of (4.5) and show that the difference is non-negative. After cross-multiplying and simplifying, we obtain a fraction with positive denominator (since each denominator in the right side of (4.5) is positive), and with numerator equal to

$$\begin{aligned} &m^2(10n_1^4 - 44n_1^3 + 12n_1^2 r + 30n_1^2 - 8n_1 r - 4n_1) \\ &+ m(15n_1^5 - 77n_1^4 + 38n_1^3 r + 74n_1^3 - 54n_1^2 r - 14n_1^2 + 8n_1 r^2 + 8n_1 r) \\ &+ (6n_1^6 - 35n_1^5 + 22n_1^4 r + 45n_1^4 - 48n_1^3 r - 12n_1^3 + 12n_1^2 r^2 + 14n_1^2 r - 4n_1 r^2). \end{aligned} \tag{4.6}$$

This expression increases with n_1 and is clearly positive when $n_1 = 6$ (to see it quickly just compare, in each parenthesis, every (maximal) sequence of consecutive negative terms with the (maximal) sequence of positive terms preceding it). Further, a direct calculation ensures that (4.6) is actually positive even when $n_1 = 5$.

However, if $n_1 \in \{3, 4\}$, then (4.6) could take on negative values for certain values of m. To deal with these two cases we revert back to the initial equation (4.4).

Assume that $n_1 = 3$. Then subtracting $\frac{n_1 - 1}{4n_1 - 2}$ from both sides of (4.4) yields that $C_1(v) - \frac{n_1 - 1}{4n_1 - 2}$ is at least

$$\frac{m+3}{2m+3} - \frac{3r}{7m+9+2r} - \frac{9-3r}{7m+15+2r} - \frac{r(m+3-r)}{10m+15+2r} - \frac{(3-r)(m-r)}{10m+21+2r} - \frac{1}{5}. \tag{4.7}$$

Placing (4.7) under one (positive) denominator, the numerator becomes

$$\begin{aligned} 1540m^4 + 2m^3(9075 - 1016r + 588r^2) + 6m^2(10605 - 1047r + 937r^2 + 112r^3) \\ + m(88155 - 3816r + 9828r^2 + 2408r^3 + 96r^4) \\ + (42525 + 1350r + 6174r^2 + 2280r^3 + 184r^4), \end{aligned} \tag{4.8}$$

which is clearly positive as $r \leqslant n_1 - 1 = 2$.

A similar calculation yields the conclusion when $n_1 = 4$. In this case, the difference of (4.4) and $\frac{n_1-1}{4n_1-2}$ yields that $C_1(v) - \frac{n_1-1}{4n_1-2}$ is at least

$$\frac{m+4}{2m+4} - \frac{2r}{5m+10+r} - \frac{8-2r}{5m+14+r} - \frac{r(m+4-r)}{14m+32+2r} - \frac{(4-r)(m-r)}{14m+40+2r} - \frac{3}{14},$$

whose numerator, when placed under a common (positive) denominator, is

$$\begin{aligned} 1855m^4 + 4m^3(5855 - 82r + 100r^2) + 2m^2(52090 + 206r + 1405r^2 + 80r^3) \\ + 4m(49180 + 2022r + 1793r^2 + 194r^3 + 4r^4) \\ + 3(44800 + 4080r + 2204r^2 + 332r^3 + 13r^4). \end{aligned}$$

This is non-negative as $r \leqslant n_1 - 1 = 3$. This concludes the proof. □

It remains to demonstrate that (P2) holds. To this end, we consider the tree G to be rooted at u and, for a node x, we let T_x be the subtree of G rooted at x. To avoid unnecessary notation later, let us observe immediately that if $\deg_G(u) = 1$ then (P2) holds. For otherwise, $n_1 \geqslant 2$ and there exists a node u' at distance two from u such that $\deg_G(u') \geqslant 2$. As a result, $W_G(u) \geqslant W_G(u') + |V(T_{u'})| - 1 > W_G(u')$, which implies that $C_1(u'; G) > C_1(u; G)$, a contradiction.

We also note that if $d_G(u, x) \leqslant 2$ for all $x \in V(G)$, then (P2) is satisfied. So assume that there exists some child of u whose subtree has depth at least 2. Among all such children of u, let z be such that $|V(T_z)|$ is maximum, that is,

$$|V(T_z)| = \max\{|V(T_v)| \,:\, v \text{ child of } u \text{ and } T_v \text{ has depth at least } 2\}.$$

We now give some notations, which are illustrated in Figure 4.3. Let $y_1, \ldots, y_t$ be the nodes of T_z with depth 2 and set $Y := \cup_{i=1}^{t} V(T_{y_i})$. Note that, by definition, $t \geqslant 1$ and $d_G(u, y_i) = 3$ whenever $1 \leqslant i \leqslant t$. Let $p_1, \ldots, p_\ell$ be the children of z (in T_z) with degree more than 1 and set $P := \{p_1, \ldots, p_\ell\}$. Let P' be the set of children of z with degree 1 and set $k := |P'|$.

Note that for any $w \in N(u)$, the definition of z ensures that T_w is a star whenever $|V(T_w)| > |V(T_z)|$. The network G' is obtained from G as follows. (An illustration is given in Figure 4.4.) For convenience, we set $n := n_0 + n_1 = |V(G)|$.

(a). For each $i \in \{1, \ldots, t\}$, the edge uy_i is added.

(b). For each $i \in \{1, \ldots, \ell\}$, the edge zp_i is removed and all other edges incident to p_i but one are removed.

(c). If there exists a child w of u different from z with $|V(T_w)| \geqslant n/2$, then we select an arbitrary set $S \subset V(T_w) \setminus \{w\}$ of size $|V(T_w)| - \lfloor n/2 \rfloor$ and, for each $s \in S$, we replace the edge sw by the edge sz.

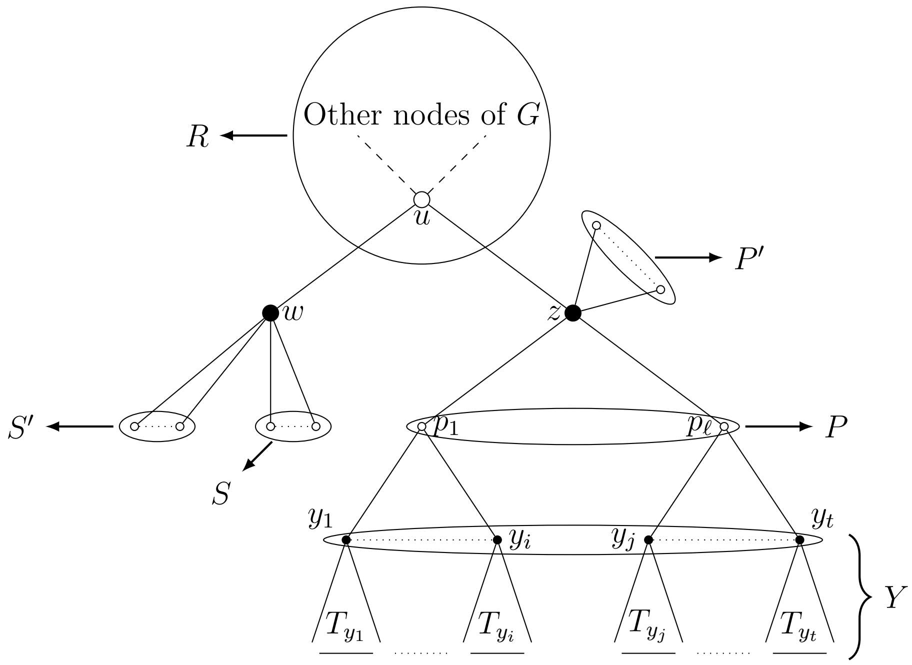


Figure 4.3: Figurative view of the subsets of nodes of G. Recall that $S' := V(T_w) \setminus w$ if $S = \varnothing$.

(d). If there is no node w as in (c), then we let w be a child of u different from z such that $|V(T_w)|$ is as large as possible, and we define S' to be $V(T_w) \setminus \{w\}$. (Recall that $\deg_G(u) \geqslant 2$, hence such a child always exists.) Moreover, we set $S := \varnothing$ for convenience.

As noted earlier, if (c) applies then T_w is a star. Moreover, if $S \neq \varnothing$, then one can see that $W_G(w) < W_G(u)$ and hence $C_1(w; G) > C_1(u; G)$. However, this is not a contradiction since $C_1(u; G) = \max\{C_1(v; G) : v \in A_0\}$ and $w \in A_1$.

Regardless of whether (c) or (d) applies, it always holds that $|S'| \leqslant \lfloor \frac{n}{2} \rfloor - 1$. Actually, it is important to notice that, in G', no child of u different from z has more than $\lfloor n/2 \rfloor - 1$ children itself. Even more, for any such child x it holds that $|V(T_x)| \leqslant \lfloor n/2 \rfloor$. This follows from our previous remark if T_x has depth at most 2, and from the fact that $|V(T_x)| \leqslant |V(T_z)|$ otherwise. Also observe that for every node $p_i \in P$, we have

$$d_G(p_i, x) = \begin{cases} d_G(u, x) - 2 & \text{if } x \in V(T_{p_i}) \\ d_G(u, x) + 2 & \text{if } x \in R \cup V(T_w) \\ d_G(u, x) & \text{otherwise.} \end{cases}$$

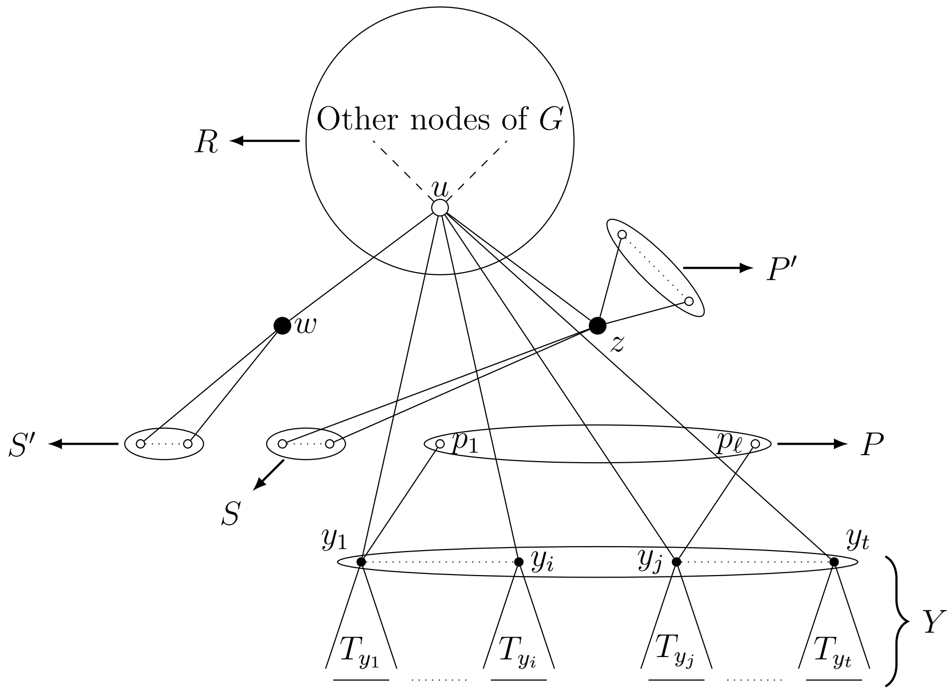


Figure 4.4: Obtaining G' from G. Recall that $S' := V(T_w) \setminus w$ if $S = \varnothing$.

Therefore, $W(p_i) \leqslant W(u) - 2(\left|V(T_{p_i})\right| - (|R| + |V(T_w)|))$. Since the definition of u implies that $W(p_i) \geqslant W(u)$, it follows that the size of $V(T_{p_i})$ is at most $\lfloor n/2 \rfloor$.

Note that G' is a tree, which we see rooted at u, and G and G' have the same node set, which we call V. In addition, G and G' have the same bipartition (A_0, A_1). Our next task is to compare the total distance of nodes in G and in G', that is, we compare $W_G(x)$ and $W_{G'}(x)$. For readability purposes, let us set $W(x) := W_G(x)$, $W'(x) := W_{G'}(x)$, and let T'_x be the subtree of G' rooted at x. We now make a few statements about $W(x)$ and $W'(x)$ for various nodes. Set $R := V \setminus V(T_z) \cup V(T_w)$ and $S' := V(T_w) \setminus (S \cup \{w\})$. We shall often use that

$$n = |V| = |R| + |Y| + |P| + \left|P'\right| + |S| + \left|S'\right| + 2.$$

Lemma 4.5. *The following hold:*

(i). If $x \in R$, then $W(x) - W'(x) = 2\,|Y|$.

(ii). If $x \in \{z\} \cup P'$, then $W'(x) \geqslant W(x) - 2\,|S|$.

(iii). If $x \in \{w\} \cup S'$, then $W'(x) = W(x) + 2\,|S| - 2\,|Y|$.

(iv). If $x \in P \cup S$, then $W'(x) \geqslant W(x)$.

(v). If $S \neq \varnothing$, then whenever $x_1 \in P'$ and $x_2 \in S'$ it holds that $W(x_1) > W(x_2)$ and $W'(x_1) > W'(x_2)$.

(vi). If $x \in Y$, then $W'(x) \leqslant W(x)$.

(vii). $W'(x) \geqslant W'(u)$ for every node $x \in Y \cup R \cup S' \cup \{w\}$.

Proof. We prove all the statements in order.

(i). If $x \in R$, then the distance from x to any node not in Y is unchanged. In addition, $d_{G'}(x, y) = d_G(x, y) - 2$ whenever $y \in Y$, hence the conclusion.

(ii). If $x \in \{z\} \cup P'$, then $d_{G'}(x, v) \geqslant d_G(x, v)$ for each $v \in V \setminus S$. In addition, if $s \in S$, then $d_{G'}(x, s) = d_G(x, s) - 2$, which yields the conclusion.

(iii). It suffices to observe that if $x \in \{w\} \cup S'$, then

$$d_{G'}(x, v) = \begin{cases} d_G(x, v) & \text{if } v \in V \setminus (S \cup Y) \\ d_G(x, v) - 2 & \text{if } v \in Y \\ d_G(x, v) + 2 & \text{if } v \in S. \end{cases}$$

(iv). First note that if $x \in P$, then the definition of G' ensures that $d_{G'}(x, v) \geqslant d_G(x, v)$ for each $v \in V$, which implies that $W'(x) \geqslant W(x)$.

Now let $x \in S$. Observe that if $v \in V$, then $d_{G'}(x, v) \geqslant d_G(x, v) - 2$. In addition, if $v \in S' \cup \{w\}$, then $d_{G'}(x, v) = d_G(x, v) + 2$. Consequently,

$$W'(x) - W(x) \geqslant 2\left|S' \cup \{w\}\right| - 2\left|V \setminus (\{x, w\} \cup S')\right|,$$

which is non-negative since $|S' \cup \{w\}| = \lfloor |V|/2 \rfloor$ when $S \neq \varnothing$, and $x \notin S' \cup \{w\}$.

(v). Let $x_1 \in P'$ and $x_2 \in S'$. Then

$$W(x_1) = 1 + 2(\left|P'\right| - 1) + \sum_{v \in P \cup Y} d_G(x_1, v) + \sum_{r \in R} d_G(x_1, r) + 3 + 4(|S| + \left|S'\right|)$$

and

$$W(x_2) = 1 + 2(|S| + \left|S'\right| - 1) + \sum_{v \in P \cup Y} d_G(x_2, v) + \sum_{r \in R} d_G(x_2, r) + 3 + 4\left|P'\right|.$$

Now, notice that $d_G(x_1, u) = 2 = d_G(x_2, u)$, which yields that

$$\sum_{r \in R} d_G(x_1, r) = \sum_{r \in R} d_G(x_2, r).$$

Similarly, for each $i \in \{1, \ldots, k\}$, it holds that $d_G(x_2, p_i) = 4 = d_G(x_1, p_i) + 2$. Therefore,

$$\sum_{v \in P \cup Y} d_G(x_2, v) = \sum_{v \in P \cup Y} d_G(x_1, v) + 2(|P| + |Y|).$$

Consequently, $W(x_1) - W(x_2)$ equals

$$\begin{aligned} & 2(|P'| - 1) + 4(|S| + |S'|) - 2(|S| + |S'| - 1) - 4|P'| - 2(|P| + |Y|) \\ = & 2(|S| + |S'| - |P| - |P'| - |Y|). \end{aligned}$$

This quantity is positive since, as $S \neq \varnothing$, we know that $|S| + |S'| \geqslant \lfloor n/2 \rfloor - 1$ while $|P| + |P'| + |Y| \leqslant n - |S| - |S'| - 3 < \lfloor n/2 \rfloor - 2$.

A similar analysis in G' yields that

$$W'(x_1) - W'(x_2) = 2(|S'| - |S| - |P'|),$$

which is again positive since $|S'| = \lfloor n/2 \rfloor - 1$ while $|P'| + |S'| \leqslant n - |S'| - 3 \leqslant \lfloor n/2 \rfloor - 2$.

(vi). Let $x \in Y$. Observe that if $d_{G'}(x, v) > d_G(x, v)$, then v must be the child of z that is an ancestor of x (that is, $v \in P$ and $x \in V(T_v)$). Furthermore, in this instance, the distance increases by exactly 2. As the distance from x to any node in R decreases by 2 (and $|R| \geqslant 1$), it follows that $W'(x) \leqslant W(x)$.

(vii). For readability, the proof is split intro three cases depending on whether $x \in \{w\}$, $x \in R$, $x \in S'$ or $x \in Y$. The interested reader will notice that a similar argument is used in all these cases, however, proceeding with cases simplifies the verification and gives a better vision of the situation.

We start by showing that $W'(w) \geqslant W'(u)$. Since $d_{G'}(w, u) = 1$, we know that

$$d_{G'}(w, v) = \begin{cases} d_{G'}(u, v) - 1 & \text{if } v \in V(T_w) \setminus S = S' \cup \{w\} \\ d_{G'}(u, v) + 1 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} W'(w) - W'(u) &= |V \setminus (S' \cup \{w\})| - |S' \cup \{w\}| \\ &= |V| - 2(|S'| + 1), \end{aligned}$$

which is non-negative since $|S'| \leqslant \lfloor n/2 \rfloor - 1$.

A similar reasoning applies to the nodes in R. Let $x \in R \setminus \{u\}$. Set $d := d_{G'}(x,u)$ and let x' be the child of u on the unique path between u and x in G. Note that $T'_{x'} = T_{x'}$. Since

$$d_{G'}(x,v) = d_{G'}(u,v) + d \qquad \text{if } v \in V \setminus V(T_{x'})$$

and

$$d_{G'}(x,v) \geqslant d_{G'}(u,v) - d \qquad \text{if } v \in V(T_{x'}),$$

we observe that

$$W'(x) - W'(u) \geqslant d \cdot (|V \setminus V(T_{x'})| - |V(T_{x'})|)\,.$$

This yields the desired inequality since, as reported earlier, $|V(T_{x'})| \leqslant n/2$.

We now deal with the nodes in S'. Let $x \in S'$. First, if $S \neq \varnothing$, then S' is composed of precisely $\lfloor n/2 \rfloor - 1$ nodes, which are all children of w. The definition of G' thus implies that $d_{G'}(x,v) \geqslant d_{G'}(u,v)$ whenever $v \neq x$, hence $W'(x) \geqslant W'(u)$, as asserted. Assume now that $S = \varnothing$. The situation can then be dealt with in the very same way as for the nodes in R. Indeed, in this case,

$$W'(x) - W'(u) \geqslant d_{G'}(x,u) \cdot (|V \setminus V(T_w)| - |V(T_w)|)\,,$$

and T_w contains at most $n/2$ nodes since $S = \varnothing$.

Finally, let $x \in Y$. Similarly as before, set $d := d_{G'}(x,u)$. For every $v \in V$, it holds that

$$d_{G'}(x,v) \geqslant d_{G'}(u,v) - d.$$

Let y_i be the ancestor of x among $\{y_1, \ldots, y_t\}$. If $v \notin V(T'_{y_i})$, then

$$d_{G'}(x,v) = d_{G'}(u,v) + d.$$

Consequently,

$$W'(x) - W'(u) \geqslant d \cdot \left(\left| V \setminus V(T'_{y_i}) \right| - \left| V(T'_{y_i}) \right| \right).$$

Now let p_k be the father of y_i in G. Then $V(T'_{y_i}) \subseteq V(T_{p_k})$. As reported earlier, $\left| V(T_{p_k}) \right| \leqslant \lfloor n/2 \rfloor$, which yields that $W'(x) - W'(u) \geqslant 0$. □

The next lemma in particular bounds $C_1(u;G)$ from below.

Lemma 4.6. *If* $x \in Y$, *then* $0 \leqslant \frac{W(x)-W'(x)}{W(x)} < 2C_1(u;G)$.

Proof. Assume that $x \in V(T_{y_i})$. Lemma 4.5(vi) ensures that $W'(x) \leqslant W(x)$, thereby proving that $\frac{W(x)-W'(x)}{W(x)}$ is non-negative.

Let D be the set of those nodes whose distance to x is greater in G than in G', that is, $D := \{v \in V : d_G(v,x) > d_{G'}(v,x)\}$. Observe that $W(x) - W'(x) \leqslant 2\,|D|$, since $d_{G'}(x,v) \geqslant d_G(x,v) - 2$ for every $v \in V$.

We partition D into parts $D_1, \ldots, D_m$ where $v \in D_j$ if and only if $v \in D$ and $d_G(x,v) = j$. Note that $D_1 = \varnothing = D_2$. In addition, $D_3 = \{u\}$ if $x \in \{y_1, \ldots, y_t\}$ while $D_3 = \varnothing$ if $x \in Y \setminus \{y_1, \ldots, y_t\}$. Finally, if $x \notin \{y_1, \ldots, y_t\}$, then $D_4 \subseteq \{u\}$, while otherwise D_4 is contained in $A_1 \setminus \{x, z\}$. In both cases, we deduce that $|D_4| \leqslant n_1 - 2$, since $n_1 \geqslant 3$. Thus

$$W(x) - W'(x) \leqslant 2 \sum_{i=3}^{m} |D_i| \tag{4.9}$$

and, since G contains at least one node at distance two from x,

$$W(x) \geqslant 1 + 2 + \sum_{i=3}^{m} i\,|D_i|\,. \tag{4.10}$$

Since we assume that $C_1(u;G) \geqslant C_1(v;H(v;n_0,n_1))$, it follows from Lemma 4.4 that $C_1(u;G) \geqslant \frac{n_1-1}{2(2n_1-1)}$. Therefore,

$$\begin{aligned}
\frac{W(x) - W'(x)}{W(x)} - 2C_1(u;G) &\leqslant \frac{W(x) - W'(x)}{W(x)} - \frac{n_1 - 1}{2n_1 - 1} \\
&\leqslant \frac{2\sum_{i=3}^{m} |D_i|}{W(x)} - \frac{n_1 - 1}{2n_1 - 1} \\
&\leqslant \frac{2(2n_1-1)\sum_{i=3}^{m} |D_i| - (n_1 - 1)(3 + \sum_{i=3}^{m} i\,|D_i|)}{(2n_1 - 1)W(x)} \\
&= \frac{-3n_1 + 3 + \sum_{i=3}^{m} |D_i|\,(n_1(4-i) - 2 + i)}{(2n_1 - 1)W(x)} \\
&\leqslant \frac{-3n_1 + 3 + |D_3|\,(n_1 + 1) + 2 \cdot |D_4|}{(2n_1 - 1)W(x)} \\
&\leqslant \frac{-3n_1 + 3 + (n_1 + 1) + 2(n_1 - 2)}{(2n_1 - 1)W(x)} \\
&= 0,
\end{aligned}$$

where the second line follows from (4.9), the third line from (4.10), and the fifth and seventh lines from our assumption that $n_1 \geqslant 3$. □

To complete the proof of Theorem 4.3, what remains is to show that $C_1(u;G') > C_1(u;G)$ which contradicts the choice of (G,u). We define

$$\gamma := \sum_{u \in \{w\} \cup S'} \frac{2\,|S|}{W(u)W'(u)} - \sum_{u \in \{z\} \cup P'} \frac{2\,|S|}{W(u)W'(u)}.$$

By Lemma 4.5(v) and the fact that $|S' \cup \{w\}| \geqslant |P' \cup \{z\}|$ whenever $S \neq \emptyset$, we infer that γ is always non-negative (noticing that $\gamma = 0$ if $S = \emptyset$).

Note that

$$C_1(u;G') - C_1(u;G) = \sum_{v \in V} \frac{1}{W'(u)} - \frac{1}{W(u)} - \left(\frac{1}{W'(v)} - \frac{1}{W(v)}\right)$$
$$= \sum_{v \in V} \frac{W(u) - W'(u)}{W(u)W'(u)} - \frac{W(v) - W'(v)}{W(v)W'(v)}.$$

For readability, set $f(v) := \frac{W(u)-W'(u)}{W(u)W'(u)} - \frac{W(v)-W'(v)}{W(v)W'(v)}$ and $g(v) := \frac{1}{W(v)W'(v)}$ for each node $v \in V$.

By Lemma 4.5(i) and (iii),

$$f(v) = \begin{cases} 2|Y|(g(u) - g(v)) & \text{if } v \in R \\ 2|Y|(g(u) - g(v)) + 2|S|g(v) & \text{if } v \in S' \cup \{w\}. \end{cases}$$

In addition, if $v \in P \cup S$ then $W'(v) \geqslant W(v)$, by Lemma 4.5(iv), so $f(v) \geqslant 2|Y|g(u)$. In total, we infer that $C_1(u;G') - C_1(u;G)$ is at least

$$\sum_{v \in Y \cup (\{z\} \cup P')} f(v) + \sum_{v \in R \cup S' \cup \{w\}} 2|Y| \cdot (g(u) - g(v))$$
$$+ 2|Y| \sum_{v \in P \cup S} g(u) + \sum_{v \in S' \cup \{w\}} 2|S| \cdot g(v).$$

Notice that $g(u) > \frac{1}{W'(u)}\left(\frac{1}{W(u)} - \frac{1}{W(v)}\right)$ for every node $v \in V$. Moreover by Lemma 4.5(i), (vi), (vii) and Lemma 4.6 we know that

$$\sum_{v \in Y} f(v) = 2|Y| \sum_{v \in Y} g(u) - \sum_{v \in Y} (W(v) - W'(v))g(v)$$
$$\geqslant 2|Y| \sum_{v \in Y} g(u) - \frac{1}{W'(u)} \sum_{v \in Y} \frac{W(v) - W'(v)}{W(v)}$$
$$> 2|Y| \sum_{v \in Y} g(u) - \frac{|Y|}{W'(u)} \cdot 2C_1(u;G)$$
$$> \frac{2|Y|}{W'(u)} \sum_{v \in Y} \left(\frac{1}{W(u)} - \frac{1}{W(v)}\right) - \frac{2|Y|C_1(u;G)}{W'(u)}.$$

So we infer that $C_1(u;G') - C_1(u;G)$ is greater than

$$\sum_{v \in P' \cup \{z\}} f(v) + 2|Y| \sum_{v \in R \cup S' \cup \{w\}} (g(u) - g(v))$$
$$+ \frac{2|Y|}{W'(u)} \sum_{v \in Y \cup P \cup S} \left(\frac{1}{W(u)} - \frac{1}{W(v)}\right) + 2|S| \sum_{v \in \{w\} \cup S'} g(v) - 2|Y| \frac{C_1(u;G)}{W'(u)}.$$

According to Lemma 4.5(vii), if $v \in R \cup S' \cup \{w\}$ then

$$g(u) - g(v) \geqslant \frac{1}{W'(u)} \left(\frac{1}{W(u)} - \frac{1}{W(v)} \right).$$

In addition, by Lemma 4.5(ii) if $v \in P' \cup \{z\}$, then

$$f(v) \geqslant 2\,|Y|\, g(u) - 2\,|S|\, g(v) > \frac{2\,|Y|}{W'(u)} \left(\frac{1}{W(u)} - \frac{1}{W(v)} \right) - 2\,|S|\, g(v).$$

Consequently, we deduce that

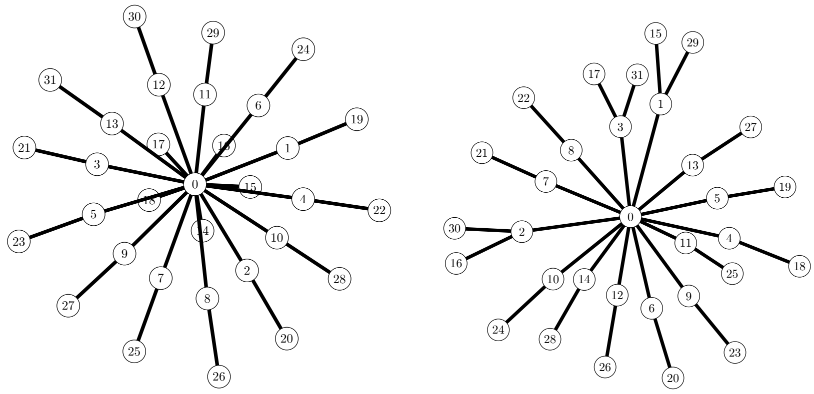
$$\begin{aligned} C_1(u;G) - C_1(u;G') &> \frac{2\,|Y|}{W'(u)} \sum_{v \in V} \left(\frac{1}{W(u)} - \frac{1}{W(v)} \right) - \frac{2\,|Y|}{W'(u)} C_1(u;G) + \gamma \\ &\geqslant \frac{2\,|Y|}{W'(u)} (C_1(u;G) - C_1(u;G)) \\ &= 0. \end{aligned}$$

This completes the proof.

4.3 CONCLUDING REMARKS AND FUTURE WORK

On Figure 4.1 we have a bipartite network N on 89 edges with partition sizes $|P_1| = 18$ and $|P_2| = 14$ that maximizes closeness centralization at nodes corresponding to Mrs. Evelyn Jefferson and to the event from September 16th, 1936 (Old City), respectively. Their closeness values are approximately equal to 0.0167 and 0.0192, while their closeness centralization values are approximately equal to 0.078 and 0.160, respectively. As shown above, the graphs $H(0, 18, 14)$ and $H(0, 14, 18)$ maximize closeness centralization among all bipartite graphs with partition sizes 11 and 28 (regarding from which partition we are measuring). These graphs are depicted on Figure 4.5. In both graphs the maximum closeness centralization is attained at the node labeled 0 with values $C_1(H(0, 14, 18), 0) \approx 0.329$ and $C_1(H(0, 11, 28), 0) \approx 0.299$, respectively.

We showed that among all two-mode networks with fixed size bipartitions n_0 and n_1, the largest closeness centralization is achieved by a rooted tree of depth 2, where neighbors of the root have an equal or almost equal number of children, namely at node v of a graph $H(v, n_0, n_1)$. This confirms a conjecture by Everett, Sinclair, and Dankelmann [48] regarding the problem of maximizing closeness centralization in two-mode data, where the number of data of each type is fixed. A similar statement for the centrality measure of eccentricity was recently established and is described in Chapter 5. However, the same conjecture remains open for eigenvector centrality.



(a) $H(0, 14, 18)$ (b) $H(0, 18, 14)$

Figure 4.5: The two graphs that maximize closeness centralization among all bipartite graphs with partition sizes 14 and 18. Note that in both cases the root is node 0.

Conjecture 4.7. *Let $\mathcal{B}(n_0, n_1)$ be the class of all bipartite graphs with bipartition P_0 and P_1 such that $|P_i| = n_i$ for $i \in \{0, 1\}$. Then*

$$\max_{G \in \mathcal{B}(n_0,n_1)} \max_{v \in P_0} \operatorname{Eig}_1(v, G) = \operatorname{Eig}_1\left(v, H\left(v, n_0, n_1\right)\right).$$

A centrality measure $\mathcal{C}$ is said to satisfy the *max-degree property* in the family $\mathcal{F}$ if for every graph $G \in \mathcal{F}$ and any node $v \in V(G)$, it holds that

$$\mathcal{C}_G(v) = \max_{u \in V(G)} \mathcal{C}_G(u) \implies \deg_G(v) = \max_{u \in V(G)} \deg_G(u).$$

While degree centrality trivially satisfies the max-degree property in $\mathcal{G}_n$, one can easily observe that this is not true for closeness centrality. Still, it is interesting to observe that the maximizing family for bipartite graphs $H\left(v, |P_0|, |P_1|\right)$ (or stars, for connected graphs $\mathcal{G}_n$ in general) both satisfies the max-degree property. It may be interesting to seek for necessary or sufficient conditions on the family $\mathcal{F}$, where closeness satisfies the max-degree property.

5

ECCENTRICITY OF NETWORKS WITH STRUCTURAL CONSTRAINTS

The concept of centrality is of fundamental importance in social network analysis. Its goal is to provide a measure of the relative "importance" of a node in a network and different motivations lead to different centrality measures. Many of the centrality concepts were first developed in social network analysis, and many of the terms used to measure centrality reflect their sociological origin [112, 51]. For further discussion on centrality, please refer to Section 3.3 on page 21.

Arguably, the most common branch of centrality functions is based on the distance between the nodes of the network. The aim of the eccentricity is to determine a node that minimizes the maximum distance to any other node in the graph.

The field of eccentricity centrality is rich with many possible applications. Recently, average eccentricity in various types of graphs has been studied by Dankelmann and others [35, 42, 70] where some computer generated and other conjectures has been settled. Interested reader can find various applications of eccentricity in [54, 55, 36, 68, 71]. Eccentricity is also studied in the field of networks (see [8]). Let us also mention, that some topological indices that involves eccentricity are used in chemical graph theory [146].

5.1 BASIC NOTIONS

The eccentricity $e_G(v)$ of a node $v \in V(G)$ in a connected network G is the maximum distance (in the network) between v and u, over all nodes u of G. For a disconnected network, all nodes are defined to have infinite eccentricity. To state this formally:

$$e_G(v) := \max\left\{d_G(v,u) \,:\, (u,v) \in V(G)^2\right\} \in \mathbb{N} \cup \{\infty\}.$$

The *center* (or *Jordan center* [138]) of a network is the set of all nodes of minimum eccentricity, that is, the set of all those nodes v such that the greatest distance $d_G(v,u)$ to other nodes u is minimal [101]. Equivalently, it is the set of nodes with eccentricity equal to the network's radius. Thus nodes in the center (also called *central* or *median points*) minimize the maximal distance from other points in the network and we define $m(G)$ to be the set of central nodes.

Based on eccentricity, Hage and Harary [63] proposed a corresponding centrality measure, namely

$$\forall v \in V(G), \quad E_G(v) := \frac{1}{e_G(v)}.$$

The reciprocal of the eccentricity value is convenient, since it obeys the *rule of monotonicity* (see Definition 3.2 on page 22).

The *centralization* of a network is a measure of how central its most central node is in relation to how central all the other nodes are. The general definition of centralization for non-weighted networks was proposed by Freeman [52] in 1979. Centralization measures then calculate the sum of differences in centrality between the most central node in a network and all other nodes; and divide this quantity by the theoretically largest such sum of differences in any network of the same degree [52]. Thus every centrality measure can have its own centralization measure. In 2006, Butts [28] studied bounds for degree centralization in graphs with different densities.

A centralization measure for eccentricity is the *eccentricity centralization*, given by

$$\forall v \in V(G), \quad E_1(G,v) := \sum_{u \in V(G)} (E_G(v) - E_G(u)).$$

When there is no risk of confusion regarding the network G, we shall write $E_1(v)$ instead of $E_1(G,v)$. We also set

$$E_1(G) := \{\max E_1(u) \; : \; u \in V(G)\}.$$

Note that $E_1(G) \geqslant 0$ for every network G. Moreover, if G is a disconnected network, then $E_1(G) = 0$ since $e_G(v) = \infty$ for every node v of G.

Freeman [52] showed that the centralizations for degree centrality, betweenness centrality and closeness centrality attain their maximum if and only if G is the star network. We provide the same result for E_1, the eccentricity centralization (see Proposition 5.5 on page 60).

If $\mathscr{C}$ is a collection of networks, we set

$$E_1^*(\mathscr{C}) := \max \{E_1(G) \; : \; G \in \mathscr{C}\}.$$

We define $\mathscr{C}^*$ to be the set of those networks G in $\mathscr{C}$ such that $E_1(G) = E_1^*(\mathscr{C})$.

Remark 5.1. For network G with n nodes, then $E_1(G,v) = E_1(G)$ if and only if $v \in m(G)$.

Indeed $E_1(G,v) = n \cdot \frac{1}{e_G(v)} - \sum_{u \in V(G)} E_G(u)$ is maximized when $e_G(v)$ is minimized, forcing v to be a median.

We provide a thorough study of $E_1^*(\mathscr{C})$ for various network classes $\mathscr{C}$, defined by prescribing parameters or structure. Specifically, we focus on bipartite networks with fixed part sizes in Section 5.2 and on networks with fixed number of nodes or edges and/or maximum degree in Section 5.3. For instance, we determine $E_1^*(\mathscr{C})$ when $\mathscr{C}$ is the class of all networks with n nodes (or n edges) and provide structural information about networks with maximum eccentricity (Proposition 5.5, Corollary 5.6). We also study the class of tree networks with fixed number of nodes and fixed maximum degree (Subsection 5.4). Among all tree networks with maximum eccentricity, we characterize those with the least number of edges and provide an efficient algorithmic way of building them all. In the course of this study, we shall develop a new way of enumerating the nodes of a tree, coined S-enumerations, which might be useful in different contexts, too.

Unless specified otherwise, every rooted tree is assumed to be rooted at a central node. In all notations, the subscripts may be omitted when there is no risk of confusion. The number of nodes of a network is its *order*, while the number of edges of a network is its *size*. We define $\mathcal{G}$ to be the collection of all networks.

We end the introduction with a straightforward, but useful, observation concerning the center of a tree.

Lemma 5.2. *Let T be a tree with diameter ℓ and let $P = v_0, \ldots, v_\ell$ be a longest path of T. If ℓ is even then $m(T) = \{v_{\ell/2}\}$ and if ℓ is odd then $m(T) = \{v_{\lfloor \frac{\ell}{2} \rfloor}, v_{\lceil \frac{\ell}{2} \rceil}\}$.*

Proof. First note that $m(T)$ contains $v_{\lfloor \frac{\ell}{2} \rfloor}$ and $v_{\lceil \frac{\ell}{2} \rceil}$ (which are the same node if ℓ is even). In addition, no other node from P belongs to $m(T)$. It is therefore enough to show that $m(T)$ is a subset of every longest path of T. Assume otherwise, and let $v \in m(T) \setminus P$. Since T is connected, there exists a unique path R that connects v with P; note that R contains at least one edge. Let $\{v_i\} = V(R) \cap V(P)$. We derive a contradiction as follows:

$$\begin{aligned} e_T(v) &= \left\lceil \frac{\ell}{2} \right\rceil \\ &\geqslant \max\left(d_G(v, v_0), d_G(v, v_\ell)\right) \\ &= |E(R)| + \max(i, \ell - i) \\ &\geqslant 1 + \frac{\ell}{2}. \end{aligned}$$

□

Lemma 5.2 will often be used implicitly in what follows.

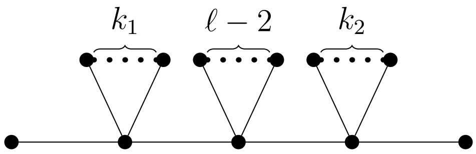


Figure 5.1: A schematic view of some members of $\mathscr{T}$, where $k_1 + k_2 = k - 3$.

5.2 BIPARTITE NETWORKS WITH FIXED PARTITION SIZES

Centrality in bipartite networks with fixed partition sizes was studied by Everett *et al.* [48] and next by Sinclair [128, 129]. They considered betweenness centralization and closeness. Following them, let us consider, given two integers k and ℓ, the family $\mathscr{C}$ of all connected networks whose nodes can be partitioned into two sets K and L such that no two nodes in K are adjacent, no two nodes in L are adjacent, $|K| = k$ and $|L| = \ell$. In this section, we focus on the notion of eccentricity centralization over the same family $\mathscr{C}$ of networks. Without loss of generality, we shall always assume that K is the larger of the two parts, that is, $k \geqslant \ell$.

Let $\mathscr{T}$ be the sub-family of $\mathscr{C}$ composed of all trees with diameter 4. (Since $\mathscr{T} \subseteq \mathscr{C}$ by definition, note that all trees in $\mathscr{T}$ admit a bipartition into two parts of order k and ℓ.) A schematic view of some of the elements in $\mathscr{T}$ is given by Figure 5.1. Our goal is to establish the following statement.

Theorem 5.3. *If $\ell \geqslant 2$ and T is a tree, then $T \in \mathscr{C}^*$ if and only if $T \in \mathscr{T}$.*

Our proof of Theorem 5.3 uses the next lemma, which states there exists a tree achieving the maximum eccentricity over $\mathscr{C}$.

Lemma 5.4. *The set $\mathscr{C}^*$ contains a tree.*

Proof. Let $G \in \mathscr{C}^*$. Fix a median v of G and a breadth-first search tree T of G rooted at v. is a bipartite network with bipartition orders k and ℓ, so $T \in \mathscr{C}$. In addition, observe that $E_T(u) \leqslant E_G(u)$ for each node $u \in V(G)$ and, further, $E_T(v) = E_G(v)$. Consequently, $E_1(T, v) \geqslant E_1(G, v)$ and hence $T \in \mathscr{C}^*$, which concludes the proof. □

Lemma 5.4 permits to deal with the case where one part of the bipartition is a singleton. Indeed, if $\ell = 1$ then there is a unique tree in $\mathscr{C}$, namely the star with $k + 1$ nodes. Thus the star on $k + 1$ nodes belongs to $\mathscr{C}^*$.

Proof of Theorem 5.3. First consider a tree T in $\mathscr{T}$ and let us compute $E_1(T)$. Since T has diameter 4, the (unique, by Lemma 5.2) median v of T is at distance at most 2 from every other node of T. If $v \in K$, then

$$E_1(T) = \frac{k+\ell-1}{2} - \frac{\ell}{3} - \frac{k-1}{4}$$
$$= \frac{\ell}{6} + \frac{k-1}{4}.$$

Let ν be $\frac{\ell}{6} + \frac{k-1}{4}$. If $v \in L$, then it would follow that

$$E_1(T) = \frac{k+\ell-1}{2} - \frac{k}{3} - \frac{\ell-1}{4}$$
$$= \frac{k}{6} + \frac{\ell-1}{4},$$

which is at most ν as $k \geqslant \ell$. Consequently, $v \in K$ and $E_1^*(\mathscr{C}) \geqslant \nu$.

The second part of the proof consists in establishing the following statement, which yields the sought conclusion:

A tree T in $\mathscr{C}$ either belongs to $\mathscr{T}$ or verifies that $E_1(T) < \nu$.

Let $T \in \mathscr{C}^*$ and $v \in m(T)$. Recall that T has a bipartition with parts of orders k and ℓ. Setting $r := e_T(v)$, we know that $e_T(u) \leqslant d_T(u,v) + r$ for every $u \in V(T)$. Therefore, letting $n_i(T)$ be the number of nodes at distance precisely i from v, and n the total number of nodes of T, we deduce that

$$E_1(T) = E_1(T,v) \leqslant \frac{n-1}{r} - \sum_{i=1}^{r} \frac{n_i(T)}{r+i}. \tag{5.1}$$

If $r \geqslant 3$, then $r+1 < 2r-1$ and since $n_1(T) \geqslant 1$, we deduce that

$$E_1(T) \leqslant \frac{n-1}{r} - \frac{1}{r+1} - \frac{\ell-1}{2r-1} - \frac{k-1}{2r}$$
$$< \frac{n-1}{r} - \frac{\ell}{2r-1} - \frac{k-1}{2r}$$
$$= \ell \cdot \left(\frac{1}{r} - \frac{1}{2r-1}\right) + (k-1) \cdot \frac{1}{2r},$$

where the first inequality uses that v and its neighbor(s) belong to different parts of T.

Let $f(r,k,\ell) := \ell \cdot (\frac{1}{r} - \frac{1}{2r-1}) + (k-1) \cdot \frac{1}{2r}$. So $f(2,k,\ell) = \nu$. Observe that if $r \geqslant 3$, then $f(r-1,k,\ell) > f(r,k,\ell)$. Indeed, first $\frac{1}{r-1} - \frac{1}{2(r-1)-1} > \frac{1}{r} - \frac{1}{2r-1}$ for $r \geqslant 3$ since the mapping $x \to \frac{x-1}{x(2x-1)}$ is (strictly) decreasing for $x \in [2, \infty)$. Second, the mapping $x \to \frac{1}{2x}$ is also decreasing, which yields the observation.

As a result, $E_1(T) < f(2,k,\ell) = \nu$ unless $r = 2$, in which case $T \in \mathscr{T}$ (and $E_1(T) = \nu$, as reported earlier). □

5.3 NETWORKS WITH PRESCRIBED ORDER OR SIZE

We are interested in maximizing E_1 over the class of networks with n nodes, for a fixed integer n. Formally, let $\mathscr{G}_n$ be the collection of all networks on n nodes. We consider

$$E_1^*(\mathscr{G}_n) = \max\{E_1(G, v) : G \in \mathscr{G}_n \text{ and } v \in V(G)\}.$$

We show that this maximum is equal to $\frac{n-1}{2}$ and is realized by networks that contain a unique *universal node* — that is, a node adjacent to all other nodes of the network — and this node is then a median and no other node is universal. In particular, a tree with a universal node is a star centered at that node.

Proposition 5.5. *Let n be an integer greater than 2. If $G \in \mathscr{G}_n$ and $v \in V(G)$, then $E_1(G, v) \leqslant \frac{n-1}{2}$ with equality if and only if v is the only universal node.*

Proof. Let $G \in \mathscr{G}_n$. Let $v \in m(G)$, so $E_1(G, v) = \sum_{u \in V(G)}(1/e(v) - 1/e(u))$. Letting r be the radius of G, we observe that $\frac{1}{r} \geqslant \frac{1}{e(u)} \geqslant \frac{1}{2r}$ for every node u. Thus we deduce that $E_1(G) \leqslant (n-1) \cdot (1/r - 1/2r) \leqslant (n-1)/2$, since $r \geqslant 1$. Furthermore, $E_1(G) = (n-1)/2$ if and only if $r = 1 = e(v)$ and $e(u) = 2$ for every $u \in V(G) \setminus \{u\}$. Consequently, v is the unique universal node of G, as stated. □

The analogous study for networks with prescribed size, rather than order, is quickly handled thanks to Proposition 5.5. For a positive integer m, let $\mathcal{G}_m$ be the collection of all networks with m edges.

Corollary 5.6. *If m is an integer greater than 1, then $E_1^*(\mathcal{G}_m) = \frac{m}{2}$ and $\mathcal{G}_m^* = \{S_m\}$.*

Proof. First, since the star with $m+1$ nodes belongs to $\mathcal{G}_m$, we know that $E_1^*(\mathcal{G}_m) \geqslant \frac{m}{2}$. Second, a network in $\mathcal{G}_m^*$ is necessarily connected and, consequently, it contains at most $m+1$ nodes. We deduce that $\mathcal{G}_m^* \subseteq \cup_{i=2}^{m+1} \mathscr{G}_i^*$. Therefore Proposition 5.5 implies that $E_1^*(\mathscr{G}_m) = \frac{m}{2}$ and every network in $\mathcal{G}_m^*$ is a network of order $m+1$ and size m with a (unique) universal node. It is therefore the star S_m with $m+1$ nodes. □

5.4 TREE NETWORKS WITH PRESCRIBED ORDER AND MAXIMUM DEGREE

As shown by Everett *et al.* [48], in the class of all trees (and, actually, of all networks) with n nodes, the star S_{n-1} maximizes the eccentricity centralization. Considering this, we are interested in maximizing E_1 over the class of trees with n nodes and maximum degree Δ, for fixed positive integers n and Δ. More precisely, let $\mathscr{T}_{n,\Delta}$ be the collection of all trees with n nodes and maximum degree Δ. We assume throughout this subsection that $\Delta < n$, as otherwise $\mathscr{T}_{n,\Delta} = \varnothing$. Our goal is to study

$$E_1^*(\mathscr{T}_{n,\Delta}) = \max\{E_1(T,v) : T \in \mathscr{T}_{n,\Delta} \text{ and } v \in V(T)\}.$$

We characterize all optimal trees from $\mathscr{T}_{n,\Delta}$ and provide an efficient (algorithmic) way to build them all. We start with some preliminary remarks.

The situation is trivial for $\Delta = 2$, as the only trees with maximum degree 2 are paths. So we assume from now on that $\Delta \geqslant 3$. Moreover, there is only one tree with maximum degree Δ and $\Delta + 1$ nodes. Similarly, there is also only one tree with maximum degree Δ and $\Delta + 2$ nodes. So we assume from now on that $n \geqslant \Delta + 3$.

A tree is Δ-*regular* if every node that has not degree Δ is a leaf. If T is a rooted tree with root r, then the *depth* of a node of T is its distance to r. The *depth of* T is the maximum of the depth over all nodes of T; in other words, it is $e_T(r)$. A Δ-regular rooted tree of depth k is *full* if every node of depth less than k has degree exactly Δ. We let $F_{\Delta,k}$ be the full Δ-regular tree with depth k. In particular, $F_{\Delta,k}$ contains $\eta(\Delta,k) := 1 + \Delta\frac{(\Delta-1)^k-1}{\Delta-2}$ nodes. As explained below, it is straightforward to obtain a (possibly tight) lower bound on the radius of a tree in terms of its maximum degree and its number of nodes.

Lemma 5.7. *Let $k(n,\Delta)$ be the smallest integer k such that $\mathscr{T}_{n,\Delta}$ contains a tree with radius k. Then*

$$k(n,\Delta) = \left\lceil \log_{\Delta-1}\left((n-1)\cdot\frac{\Delta-2}{\Delta}+1\right)\right\rceil.$$

Proof. Fix $T \in \mathscr{T}_{n,\Delta}$ and let k be the radius of T. Rooting T at a median, one sees that n is at most

$$1 + \Delta + \Delta(\Delta-1) + \cdots + \Delta(\Delta-1)^{k-1} = 1 + \Delta\frac{(\Delta-1)^k - 1}{\Delta-2}.$$

So, $(\Delta-1)^k \geqslant (n-1)\cdot\frac{\Delta-2}{\Delta} + 1$, and hence

$$k \geqslant \log_{\Delta-1}\left[(n-1)\cdot\frac{\Delta-2}{\Delta}+1\right].$$

This shows that $k(n,\Delta) \geqslant \lceil \log_{\Delta-1}((n-1)(\Delta-2)/\Delta+1)\rceil$. The equality is now straightforward. □

Let T be a tree of diameter d and assume that $v_0,\ldots,v_d$ is a longest path of T. Then by Lemma 5.2 the radius of T is $k := \lceil d/2\rceil$ and every node is at distance at most k from each of $v_{\lfloor d/2\rfloor}$ and $v_{\lceil d/2\rceil}$. Consider now T to be rooted at v_k. For each $i \in \{1,\ldots,k\}$, the *layer* i of T is defined to be the set $L_i(T)$ of all nodes v of T with depth i, that is, such that $d_T(v,v_k) = i$. We set $n_i(T) := |L_i(T)|$. If uv

is an edge such that $u \in L_i(T)$ and $v \in L_{i+1}(T)$, then u is the *parent* of v and v is a *child* of u.

We shall demonstrate that, informally, every tree T in $\mathscr{T}^*_{n,\Delta}$ has diameter $2k(n,\Delta)$ and, subject to this, the following structure: $n_k(T)$ is as large as possible, while $n_i(T)$ contains, almost always, just as many vertices as needed so that every node in $n_{i+1}(T)$ can have a parent (recall that the maximum degree cannot exceed Δ), for $i \in \{1,\dots,k-1\}$.

Before being precise, let us note a straightforward fact: in any tree T rooted at its median and with even diameter, if one "re-arranges" the subtrees rooted at any fixed level so that neither the diameter nor the maximum degree changes, then the eccentricity centralization of the tree does not change either. Specifically, this follows from the fact that if a tree T with n nodes and diameter $2k$ is rooted at its median, then $E_1(T) = \frac{n-1}{k} - \sum_{i=1}^{k} \frac{n_i(T)}{k+i}$. Let $v \in L_{i+1}(T)$ and define T' as the tree obtained from T by deleting the edge between v and its parent and adding an edge between v and any node in $L_i(T)$ of degree less than Δ. If T' has the same diameter as T, then it follows that $E_1(T') = E_1(T)$. In this case, the operation is said to be *valid*. Valid operations yield an equivalence relation between trees: two trees are *equivalent* if one is obtained from the other by a sequence of valid operations.

In the next subsection, we define a class $\mathcal{F}_{\Delta,k}(n)$ of trees all having fixed order n, maximum degree Δ and (even) diameter $2k$. As we shall see, this class captures all trees with maximum eccentricity, in the sense that every tree with n nodes, maximum degree Δ and maximum eccentricity is a member of $\mathcal{F}_{\Delta,k(n,\Delta)}(n)$. To this end, we introduce S-enumerations of trees and prove a couple of useful properties of these enumerations.

5.4.1 *S-enumerations*

We give an algorithmic procedure to label the vertices of a tree on n nodes all differently with labels $0,1,\dots,n-1$. Let T be a tree with n nodes and diameter d:

- We start from a longest path of T and label its nodes consecutively with $0,\dots,d$.
- We then consecutively label only those unlabeled nodes with a labeled parent. To this end, the following loop is performed. For i from 1 to $\lfloor d/2 \rfloor$, do the following two loops, in order:
    1. For each unlabeled child v of the node labeled i, label the nodes in the subtree rooted at v according to a depth-first search algorithm.

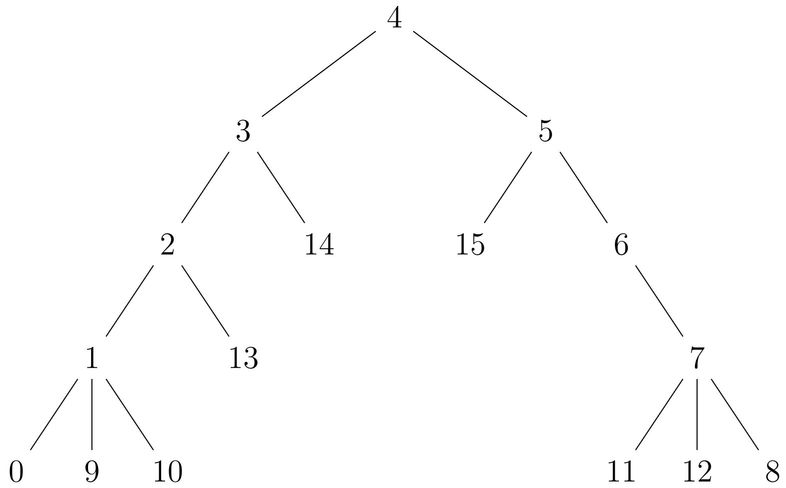


Figure 5.2: An S-enumeration of a tree with maximum degree 4, diameter 8 and 16 nodes.

2. For each unlabeled child v of the node labeled $d - i$, label the nodes at v according to a depth-first search algorithm.

Note that the running time of the procedure is $O(|V(T)|)$. Besides, the labeling is not uniquely defined, as it depends on the longest path chosen as well as the order in which the nodes are considered in each depth-first search procedure. Any labeling of the nodes of a tree T that can be obtained by the above procedure is called an *S-enumeration* of T. The longest path used in the S-enumeration is called the *root-path*. Figure 5.2 provides an example of an S-enumeration of a tree with maximum degree 4, diameter 8 and 16 nodes.

For positive integers Δ, k and n with $n \geqslant \max\{\Delta + 1, 2k\}$, let $F_{\Delta,k}(n)$ be the (unique) subtree of an S-enumeration of the full tree $F_{\Delta,k}$ induced by the nodes with labels in $\{0, \ldots, n-1\}$. Thus $F_{\Delta,k}(n)$ has n nodes, maximum degree Δ, radius k and diameter $2k$. The tree $F_{4,4}(16)$ is depicted in Figure 5.3. Let $\mathcal{F}_{\Delta,k}(n)$ be the collection of all trees that are equivalent to $F_{\Delta,k}(n)$.

The tree given in Figure 5.2 does not belong to $\mathcal{F}_{4,4}(16)$. Indeed, this tree contains two nodes on its root-path (namely 2 and 3) such that both have non-trivial subtrees, and the one further to the median is not full. In particular, consider a tree T in $\mathcal{F}_{\Delta,k}(n)$ with one of its S-enumerations: if u and v are two nodes on the root-path such that the level of u is smaller than that of v, then the degree of u cannot be greater than that of v.

We are now in a position to state the characterization of trees with maximum eccentricity.

Theorem 5.8. *Let Δ and n be integers such that $3 \leqslant \Delta \leqslant n-3$. It holds that $\mathscr{T}^*_{n,\Delta} = \mathcal{F}_{\Delta,k(n,\Delta)}(n)$.*

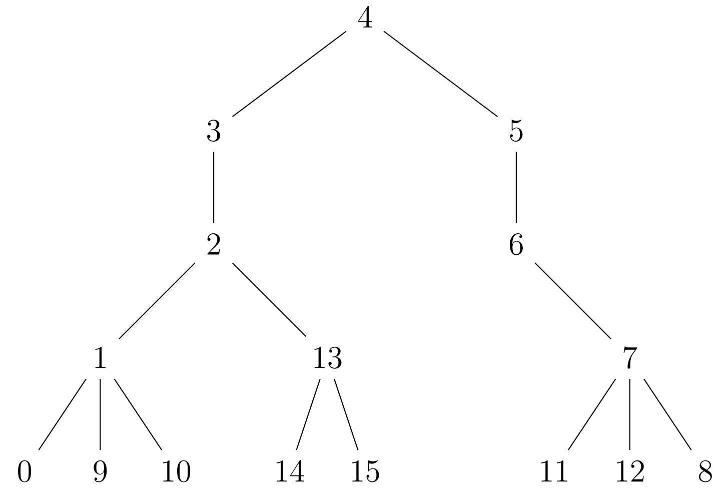


Figure 5.3: An S-enumeration of the tree $F_{4,4}(16)$.

To prove Theorem 5.8, we first establish that $F_{\Delta,k}(n)$ admits a particular partition of its nodes, which turns out to be useful to us. A path P of a tree T is *monotone* if P does not contain more than one node of each possible depth. In addition, let us define a *D-tree of depth k* to be a rooted tree in which every node of depth less than k has exactly D children. It will be useful to note that such a tree contains exactly D^i nodes of depth i for $i \in \{0, \dots, k\}$, and consequently precisely $\nu(D,k) := \frac{D^{k+1}-1}{D-1}$ nodes in total. The partition defined in the next lemma is maybe better digested when read along with the example given after the proof of that lemma.

Lemma 5.9. *Define t to be the number of leaves of $F := F_{\Delta,k}(n)$. There exists a partition of the nodes of T into t sets $V_1, \dots, V_t$ such that*

1. *for each $i \in \{1, \dots, t\}$, the nodes in V_i induce a monotone path in F;*
2. *$|V_1| = k$ and $|V_i| \in \{1, \dots, k-1\}$ if $2 \leqslant i \leqslant t$;*
3. *each set V_i contains exactly one leaf of F and, if $i < t$, then this leaf has depth k; and*
4. *for every $\ell \in \{1, \dots, k-1\}$, if F contains $(j+2)$ nodes of depth ℓ for a positive integer j, then the number of $(\Delta-1)$-trees of depth $k-\ell$ in F is at least $j+1$; moreover, $(j+1)(\Delta-1)^{k-\ell-1} \cdot (\Delta-2)$ sets among $(V_i)_{0 \leqslant i \leqslant t+1}$ have order exactly 1 and are composed of a single leaf of depth k.*

Proof. The sought partition can be built as follows. Start from an S-enumeration of $F_{\delta,k}$ such that F is the subtree induced by the nodes with labels in $\{0, \dots, n-1\}$. Let L be the set of nodes of F of depth k. Observe that $|L| \in \{t-1, t\}$. Let

$v_1, \dots, v_{|L|}$ be the elements of L increasingly ordered with respect to their label in the enumeration, so the label of v_1 is 0 and that of v_2 is $2k$. For convenience, set $V_0 := \emptyset$. For each index i from 1 up to $|L|$, let V_i be the set of nodes of the longest monotone path containing v_i in the network $F - \cup_{0 \leqslant j < i} V_j$. Observe that if $|L| = t$, that is, all leaves of F have depth k, then $(V_i)_{1 \leqslant i \leqslant t}$ is a partition of $V(F)$ into non-empty parts. If $|L| = t - 1$, then we further define the set V_t to be $V(F) \setminus \cup_{1 \leqslant j \leqslant t-1} V_j$. Note that V_t contains a leaf (of depth less than k) and induces a monotone path. Either way, $(V_i)_{1 \leqslant i \leqslant t}$ is a partition of $V(F)$ into t non-empty parts.

The partition $(V_i)_{1 \leqslant i \leqslant t}$ of the nodes of F readily satisfies properties 1, 2 and 3. It remains to prove that property 4 is satisfied. To this end, let $x_1, \dots, x_{j+2}$ be the nodes of F of depth ℓ ordered increasingly with respect to their labels, so x_1 is the node labeled $k - \ell$ and x_2 the node labeled $k + \ell$. Assume that $j \geqslant 1$. Then the definition of F implies that the subtree rooted at x_i is a $(\Delta - 1)$-tree of depth $k - \ell$ whenever $1 \leqslant i \leqslant j + 1$, which implies the first statement. Moreover, our construction of the partition implies that for each node v of depth $k - \ell - 1$ in such a tree, exactly $\Delta - 2$ children of v are contained in a part of order 1. Therefore, in total, the partition $(V_i)_{1 \leqslant i \leqslant t}$ contains at least $(j+1)(\Delta - 1)^{k-\ell-1} \cdot (\Delta - 2)$ parts of order 1. □

We use the following convention: when we build a partition of the nodes of $F_{\Delta,k}(n)$, we use the procedure given in the preceding proof and the leaves of $F_{\Delta,k}(n)$ are considered in increasing order with respect to their labels in the S-enumeration.

Example 5.10. The partition obtained for the tree depicted in Figure 5.3 is $V_0 := \{0, 1, 2, 3, 4\}$, $V_1 := \{8, 7, 6, 5\}$, $V_2 := \{9\}$, $V_3 := \{10\}$, $V_4 := \{11\}$, $V_5 := \{12\}$, $V_6 := \{14, 13\}$ and $V_7 := \{15\}$. As stated by Lemma 5.9 (because T has more than two nodes of depth 3), this partition contains at least $2(\Delta - 1)^{k-3-1} \cdot (\Delta - 2) = 4$ singletons.

Eccentricity relates to S-enumerations of full regular trees as indicated in the next lemma.

Lemma 5.11. *If T is a tree in $\mathscr{T}_{n,\Delta}$ with diameter $2k$, then $E_1(T') \geqslant E_1(T)$ for every $T' \in \mathcal{F}_{\Delta,k}(n)$.*

To prove Lemma 5.11 we first recall that (5.1) ensures that if T is a tree with n nodes and diameter $2k$ rooted at its median, then $E_1(T) = \frac{n-1}{k} - \sum_{i=1}^{k} \frac{n_i(T)}{k+i}$. Moreover, if T has maximum degree Δ, then we know that $n_1(T) \leqslant \Delta$ and $n_{i+1}(T) \leqslant n_i(T)(\Delta - 1)$ if $1 \leqslant i < k$. This motivates the introduction of the following (more general) integer program.

Definition 5.12. Let k, Δ and n be positive integers such that $\max\{2k-1,\Delta\} < n < \eta(\Delta,k)$ and $\Delta \geqslant 2$. Let $\alpha_1,\ldots,\alpha_k$ be a decreasing sequence of positive rational numbers. The integer program (P) with parameters k, Δ, n and $(\alpha_i)_{i=1}^k$ is

$$\min \quad \sum_{i=1}^{k} \alpha_i \cdot n_i \tag{5.2}$$

$$\text{s.t.} \quad \sum_{i=1}^{k} n_i = n \tag{5.3}$$

$$n_1 \leqslant \Delta \tag{5.4}$$

$$n_{i+1} \leqslant n_i(\Delta-1) \quad \text{if } i \in \{1,\ldots,k-1\} \tag{5.5}$$

$$n_i \in \mathbf{N} \setminus \{0,1\} \quad \text{if } i \in \{1,\ldots,k\} \tag{5.6}$$

It turns out that the optimal solutions of (P) can be determined and they correspond to the sizes of the layers in specific trees with n nodes, diameter $2k$ and maximum degree Δ. Our strategy to prove Theorem 5.8 is to reduce the problem to the program (P) with some well-chosen parameters. In particular, in the proof of Theorem 5.8, the program (P) will be considered with parameter $n-1$ instead of n.

We solve the program (P) in the next proposition. Recall that $\eta(\Delta,k)$ is $1+\Delta\frac{(\Delta-1)^k-1}{\Delta-2}$, the number of nodes in the full Δ-regular tree of depth k, while $\nu(D,k)$ is $\frac{D^{k+1}-1}{D-1}$, the number of nodes in the D-tree of depth k. We shall often use that $\eta(\Delta,k) = 1+\Delta\cdot\nu(\Delta-1,k-1)$.

Proposition 5.13. *Let k, Δ and n be positive integers such that $2k+\Delta-1 \leqslant n < \eta(\Delta,k)$, and $\Delta \geqslant 3$. Let $\alpha_1,\ldots,\alpha_k$ be a (strictly) decreasing sequence of positive rational numbers. The optimal value of the integer program (P) with parameters k, Δ, n and $(\alpha_i)_{i=1}^k$ is attained only by the feasible solution obtained in the following inductive way. Setting $n_0 := 0$, we define n_i, for each $i \in \{1,\ldots,k-1\}$, to be the least integer $s \geqslant 2$ such that $s\cdot\nu(\Delta-1,k-i) \geqslant n-\sum_{j=0}^{i-1} n_j$. Finally, n_k is defined to be $n-\sum_{j=0}^{k-1} n_j$.*

Proof. For convenience, we set $\sigma_i := \sum_{j=0}^{i} n_j$ for $i \in \{0,\ldots,k\}$. First, we need to prove that the obtained solution $(n_1,\ldots,n_k)$ is feasible. As a preliminary remark, we note that $n_k \geqslant 0$ since $\nu(\Delta-1,k-i) \geqslant \frac{(\Delta-1)^2-1}{\Delta-2} \geqslant 3$ whenever $1 \leqslant i \leqslant k-1$, since $\Delta \geqslant 3$. Now, notice that (5.3) is satisfied since n_k is defined to be $n-\sigma_{k-1}$. Moreover, $n_1 \leqslant \Delta$ since $\Delta\cdot\nu(\Delta-1,k-1) = \eta(\Delta,k)-1 \geqslant n$.

We now prove that (5.5) is satisfied. Let $i \in \{1, \ldots, k-1\}$. Since $n_i \cdot \nu(\Delta - 1, k-i) \geqslant n - \sigma_{i-1}$, we deduce that

$$\begin{aligned} n - \sigma_i &= (n - \sigma_{i-1}) - n_i \\ &\leqslant n_i(\nu(\Delta - 1, k-i) - 1) \\ &= n_i \left(\frac{(\Delta-1)^{k-i+1} - (\Delta - 1)}{\Delta - 2} \right) \\ &= n_i(\Delta - 1) \cdot \nu(\Delta - 1, k - (i+1)), \end{aligned}$$

and hence $n_{i+1} \leqslant n_i(\Delta - 1)$.

It remains to prove that $n_k \geqslant 2$. If $n_i = 2$ for every $i \in \{1, \ldots, k-1\}$, then $n_k \geqslant 2$ since $n \geqslant 2k$. Otherwise, let i be the largest integer such that $n_i \geqslant 3$ and let us prove that $n_k \geqslant 2$. First, the definition of n_i implies that $n \geqslant \sigma_{i-1} + (n_i - 1)\nu(\Delta - 1, k-i) + 1$. Moreover, $\sigma_{k-1} = \sigma_{i-1} + n_i + 2(k-1-i)$. Since $n_k = n - \sigma_{k-1}$, it follows that

$$n_k - 3 \geqslant n_i(\nu(\Delta - 1, k-i) - 1) - \nu(\Delta - 1, k-i) - 2k + 2i. \tag{5.7}$$

It therefore suffices to prove that

$$n_i(\nu(\Delta - 1, k-i) - 1) - \nu(\Delta - 1, k-i) - 2k + 2i \geqslant -1. \tag{5.8}$$

Since $n_i \geqslant 3$ and $\nu(\Delta - 1, k-i) \geqslant 3$, it thus suffices to prove that $2\nu(\Delta - 1, k - i) - 2(k+1-i) \geqslant 0$. This holds because $\nu(\Delta - 1, k-i) \geqslant 2^{k-i+1} - 1 \geqslant k+1-i$ as $i \in \{1, \ldots, k-1\}$ and $\Delta \geqslant 3$. Since all constraints from Definition 5.12 are verified, $(n_1, \ldots, n_k)$ is feasible.

The optimality of $(n_1, \ldots, n_k)$ follows from the fact that $(\alpha_i)_{i=0}^k$ is a (strictly) decreasing sequence of positive numbers. Let $(n'_1, \ldots, n'_k)$ be a feasible solution. Since $\sum_{i=0}^k \alpha_i \cdot n_i \leqslant \sum_{i=0}^k \alpha_i \cdot n'_i$, we may assume that $n'_i < n_i$ for some index $i \in \{1, \ldots, k\}$. Let i be the least positive integer such that $n'_i < n_i$. Observe that $i > 1$. Indeed, if $n'_1 < n_1$, then as $n'_1 \geqslant 2$ the definition of n_1 implies that $n'_1 \nu(\Delta - 1, k-1) < n$. On the other hand, since $n'_{j+1} \leqslant n'_j(\Delta - 1)$ for each $j \in \{1, \ldots, k-1\}$ by (5.5), we deduce that

$$\sum_{j=1}^{k} n'_j \leqslant n'_1 \sum_{j=1}^{k} (\Delta - 1)^{j-1} = n'_1 \nu(\Delta - 1, k-1) < n,$$

contrary to (5.4). This contradiction ensures that $i > 1$.

We assert that $n'_j = n_j$ for each $j \in \{1, \ldots, i-1\}$. Otherwise, let $\ell \in \{1, \ldots, i-1\}$ such that $n'_\ell > n_\ell$ and $n'_j = n_j$ if $\ell < j < i$. Let $\mathbf{x} = (x_1, \ldots, x_k)$ be defined by

$$x_j := \begin{cases} n'_\ell - 1 & \text{if } j = \ell \\ n'_i + 1 & \text{if } j = i \\ n'_j & \text{otherwise.} \end{cases}$$

Then $\sum_{j=1}^{k} \alpha_j \cdot x_j = \sum_{j=1}^{k} \alpha_j \cdot n'_j + (\alpha_i - \alpha_\ell) < \sum_{j=1}^{k} \alpha_j \cdot n'_j$, which shows that if $\mathbf{x}$ is feasible, then $(n'_1, \ldots, n'_k)$ is not optimal. Thus it remains to prove that $\mathbf{x}$ is feasible to conclude the proof of our assertion.

Note that $\sum_{j=1}^{k} x_j = n$ by the definition. Moreover, $x_\ell = n'_\ell - 1 \geqslant n_\ell \geqslant 2$ and $x_i > n'_i \geqslant 2$. Surely, $x_{j+1} \leqslant x_j(\Delta - 1)$ if $j \notin \{\ell, i-1\}$. It remains to prove that $x_{\ell+1} \leqslant x_\ell(\Delta - 1)$ and $x_i \leqslant x_{i-1}(\Delta - 1)$. (These two inequalities are the same if $\ell = i - 1$.) For the sake of clarity, assume first that $\ell \neq i - 1$. Then, the former inequality holds because

$$x_{\ell+1} = n'_{\ell+1} = n_{\ell+1} \leqslant n_\ell(\Delta - 1) \leqslant (n'_\ell - 1)(\Delta - 1) = x_\ell(\Delta - 1),$$

while the latter inequality holds because

$$x_i = n'_i + 1 \leqslant n_i \leqslant n_{i-1}(\Delta - 1) = n'_{i-1}(\Delta - 1) = x_{i-1}(\Delta - 1).$$

If $\ell = i - 1$, then

$$x_i = n'_i + 1 \leqslant n_i \leqslant n_{i-1}(\Delta - 1) \leqslant (n'_{i-1} - 1)(\Delta - 1) = x_{i-1}(\Delta - 1).$$

Therefore, $\mathbf{x}$ is feasible if $n'_j > n_j$ for some $j \in \{1, \ldots, i-1\}$. We conclude that $n'_j = n_j$ if $j < i$.

However, this leads to a contradiction. Indeed, since $2 \leqslant n'_i < n_i$, the definition of n_i implies that $n'_i \cdot \nu(\Delta - 1, k - i) < n - \sigma_{i-1}$. Moreover, $\sigma_{i-1} = \sum_{j=1}^{i-1} n'_j$ by what precedes. But

$$\sum_{j=i}^{k} n'_j \leqslant n'_i \cdot \sum_{j=i}^{k} (\Delta - 1)^{j-i} = n'_i \cdot \nu(\Delta - 1, k - i),$$

which implies that $\sum_{j=1}^{k} n'_j \leqslant \sigma_{i-1} + n'_i \cdot \nu(\Delta - 1, k - 1) < n$, a contradiction. □

A key consequence of Proposition 5.13 is that if a tree T belongs to $\mathcal{F}_{\Delta,k}(n)$, then the vector $(n_1(T), \ldots, n_k(T))$ is the optimal solution of the program (P) with parameters k, $n - 1$, Δ and $\alpha_i := \frac{1}{k+i}$ for $i \in \{1, \ldots, k\}$. Lemma 5.11 follows from this observation.

Proof of Lemma 5.11. Let T be a tree in $\mathscr{T}_{n,\Delta}$ with diameter $2k$. Set $n := |V(T)|$ and let $T' \in \mathcal{F}_{\Delta,k}(n)$. The vector $(n_1(T), \ldots, n_k(T))$ is a feasible solution of the program (P) with parameters k, $n - 1$, Δ and $\alpha_i := \frac{1}{k+i}$ for $i \in \{1, \ldots, k\}$. Therefore $\sum_{i=1}^{k} \frac{n_i(T)}{k+i} \geqslant \sum_{i=1}^{k} \frac{n_i(T')}{k+i}$ by the remark above. Consequently $\frac{n-1}{k} - \sum_{i=1}^{k} \frac{n_i(T)}{k+i}$ is at most $\frac{n-1}{k} - \sum_{i=1}^{k} \frac{n_i(T')}{k+i}$, which is to say that $E_1(T)$ is at most $E_1(T')$. □

We are now ready to establish Theorem 5.8.

5.4.2 *The Proof of Theorem 5.8*

Let T be a tree in $\mathscr{T}^*_{n,\Delta}$ and let d be the diameter of T. So $n \geqslant d+\Delta-1 \geqslant d+2$, as $\Delta \geqslant 3$. Clearly, the conclusion holds for networks with diameter less than 3, so we assume that d is at least 3. Our first aim is to show that the diameter d of T is $2k(n,\Delta)$. (Recall that $k(n,\Delta)$ is defined in Lemma 5.7.) We set $k_0 := k(n,\Delta)$ for convenience and proceed in two steps: we establish that d is even and, next, we prove that if $d \geqslant 2k_0+2$, then there exists a network T' of diameter $d-2$ with n nodes and maximum degree Δ such that $E_1(T') > E_1(T)$.

Suppose, for a contradiction, that $d = 2s+1$ for some positive integer s. Then the number n of nodes of T is at least $2s+3$. In addition, T must contain a longest path $P = v_0,\dots,v_d$ and a leaf that does not belong to P. Suppose first that there exists a leaf u not on P such that $T-u$ still has maximum degree Δ. Then let T' be the network obtained from T by deleting the edge incident to u and adding the edge $\{u, v_d\}$. The network T' is a tree of diameter $d+1 = 2s+2$ with n nodes and maximum degree Δ. Moreover, as $e_{T'}(v) \geqslant e_T(v)$ for every node v with strict inequality for (exactly) one of v_s and v_{s+1}, it follows that $E_1(T') > E_1(T)$, which is a contradiction. Thus we may in particular assume that T has a unique node v of degree Δ, all leaves of T not on P are adjacent to v and there are at least $\Delta-2$ of them. In addition, note that exactly one node v_i of P has degree greater than 2. Without loss of generality, we may assume that $i \leqslant s+1$.

If $v_i = v$, that is, v_i is the unique node of T with degree Δ, then $n = d+\Delta-1$ and T is composed of the path P and $\Delta-2$ leaves attached to v_i. In this case, a straightforward check ensures $E_1(T)$ is maximized only if $i = 1$ (recalling that $i \leqslant s+1$). Let T' be the tree obtained by deleting the edge incident to v_d and next adding an edge between v_d and v_{d-2}. The tree T' has n nodes and maximum degree Δ. Moreover, one sees that

$$E_1(T') = E_1(T', v_s) = \frac{n-1}{s} - 2\sum_{i=1}^{s-1}(s+i)^{-1} - \frac{\Delta+1}{2s}.$$

As

$$E_1(T) = E_1(T, v_{s+1}) = \frac{n}{s+1} - 2\sum_{i=1}^{s}(s+i)^{-1} - \frac{\Delta}{2s+1},$$

we deduce that

$$\begin{aligned} E_1(T') - E_1(T) &= \frac{n-s-1}{s(s+1)} + \frac{2s+1-\Delta}{2s(2s+1)} \\ &= \frac{s(4n-2s-3-\Delta)+2n-1-\Delta}{2s(s+1)(2s+1)} \\ &= \frac{s(3n-3)+n+2s-1}{2s(s+1)(2s+1)} > 0, \end{aligned}$$

where the last line uses that $n = d + \Delta - 1 = 2s + \Delta > 1$.

We conclude that v is not on P. In this case, v is adjacent to exactly $\Delta - 1$ leaves $u_1, \ldots, u_{\Delta-1}$. For each $i \in \{1, \ldots, \Delta - 1\}$, we delete the edge $\{v, u_i\}$ and add the edge $\{v_d, u_i\}$. It follows that T' has maximum degree Δ, order n, diameter $d+1$ and $E_1(T') > E_1(T)$, a contradiction. This contradiction shows that d must be even.

Suppose now that $d = 2k + 2$ with $k \geqslant k(n, \Delta)$. In particular, $k \geqslant 2$ since $d \geqslant 4$. Our goal is to obtain a contradiction by showing the existence of a tree T' with n nodes, maximum degree Δ and diameter $2k$ such that $E_1(T') > E_1(T)$. Since d is even, Lemma 5.11 allows us to assume that T belongs to $\mathcal{F}_{\Delta,k+1}(n)$. Recall that $n \leqslant \eta(\Delta, k(n, \Delta)) \leqslant \eta(\Delta, k)$.

Notice that the median of T, which is the node labeled $k+1$ by Lemma 5.2, has degree 2 in T. Indeed, if it had degree more than 2, then as T belongs to $\mathcal{F}_{\Delta,k+1}(n)$, we infer that the subtrees of T rooted at the nodes labeled k and $k+2$ are both $(\Delta-1)$-trees of depth k by Lemma 5.9(4). Consequently, each of these trees contains $\nu(\Delta - 1, k)$ nodes. Therefore, the total number of nodes of T would be greater than $2 \cdot \nu(\Delta - 1, k)$, which is at least $\eta(\Delta, k)$ as $\Delta \geqslant 3$ and $k \geqslant 2$, a contradiction.

In other words, T contains exactly two nodes of depth 1. A similar counting argument allows us to establish that T contains at most Δ nodes of depth 2. Indeed, let x be the number of nodes of T of depth 2, hence $2 \leqslant x \leqslant 2\Delta - 2$. Since $T \in \mathcal{F}_{\Delta,k+1}(n)$, if $x > 2$ then all but at most one of the x subtrees of T rooted at the nodes of depth 2 are $(\Delta - 1)$-trees of depth $k - 1$. Consequently, T contains more than $1 + (x-1) \cdot \nu(\Delta - 1, k - 1)$ nodes. Therefore,

$$1 + (x-1) \cdot \nu(\Delta - 1, k-1) < n \leqslant \eta(\Delta, k) = 1 + \Delta \cdot \nu(\Delta - 1, k - 1),$$

which implies that $x - 1 < \Delta$, that is, $x \leqslant \Delta$ as asserted.

We now define a new tree T'. (An example of the construction is given in Figure 5.4.) Let $P = v_0, \ldots, v_d$ be the root-path of T, that is, the path induced by nodes with labels in $\{0, \ldots, d\}$. For each node v in P, let $S(v)$ be the collection of all neighbors of v in T that do not belong to P. To obtain T', we start from a path $v'_1, \ldots, v'_{d-1}$, so, in particular, T' will have diameter at least (and, actually, exactly) $d - 2$. For each i from 1 to $d - 1$ and for each node v in $S(v_i)$, we define T_v to be the subtree of T rooted at v. We add to T' a copy of T_v and join its root to the node v'_j of T' with

$$j := \begin{cases} i+1 & \text{if } i \in \{1, \ldots, k\} \text{ and} \\ i-1 & \text{if } i \in \{k+2, \ldots, d-1\}. \end{cases}$$

Note that, as we proved earlier, the node labeled $k+1$ in T has exactly two children, which both belong to P. Hence the tree T' is well defined and so far

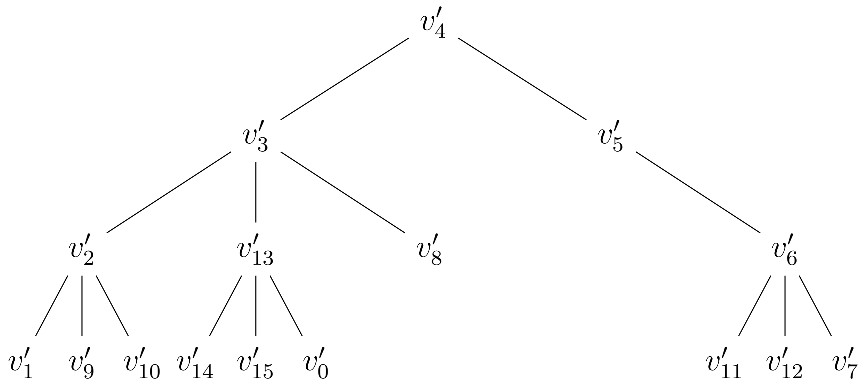


Figure 5.4: The tree T' obtained if T is $F_{4,4}(16)$, the tree of Figure 5.3.

it contains exactly $n-2$ nodes. Moreover, as proved earlier, T contains at most Δ nodes of depth 2. Consequently, the degree of $v_{d/2}$ in T' is at most Δ. In total, the maximum degree of T is hence exactly Δ. Last, the radius of T' is k since each node is at distance at most k from $v_{d/2}$ by the construction.

We finish the construction of T' by doing twice the following: among all nodes of degree less than Δ and depth less than k, we choose a node v with the largest possible depth and we add a new neighbor to v. These last two steps are always possible, since $n \leqslant 1 + \Delta\frac{(\Delta-1)^k-1}{\Delta-2}$. In case there are more than one such node, we choose the one corresponding to the node of T with the smallest label. Let v'_0 and v'_d be these two added nodes.

Observe that $T' \in \mathcal{F}_{\Delta,k}(n)$, with root-path $\{v'_1, \dots, v'_{d-1}\}$. Notice also that there is a natural one-to-one correspondence between the nodes of T and T', with v'_0 and v'_d corresponding to v_0 and v_d. Consequently, we shall make no distinction between nodes of T and T' in what follows, and we call V the common set of nodes of T and T'.

It remains to show that $E_1(T') > E_1(T)$. To this end, we set for convenience $\mu'(v) := E_{T'}(v_{k+1}) - E_{T'}(v)$ and $\mu(v) := E_T(v_{k+1}) - E_T(v)$ for every node $v \in V$. We consider partitions $(V_i)_{1\leqslant i\leqslant t}$ and $(V'_i)_{1\leqslant i\leqslant t'}$ of the nodes of T and T' given by Lemma 5.9, respectively. Notice that $t' \in \{t, t+1, t+2\}$, depending on whether v'_0 was joined to a node of degree $\Delta-1$ or not and of depth $k-1$ or less (recall that v'_0 is the last but one node added to T' in the construction process). Hence

$$E_1(T') - E_1(T) \geqslant \sum_{i=1}^{t}\left(\sum_{v\in V'_i}\mu'(v) - \sum_{v\in V_i}\mu(v)\right).$$

We shall now establish that $E_1(T') - E_1(T) > 0$ by proving that $\sum_{v\in V'_i}\mu'(v) - \sum_{v\in V_i}\mu(v) \geqslant 0$ for each $i \in \{1,\dots,t\}$, with strict inequality for at least one index.

The set V_1 is composed of $v_0, \ldots, v_{k+1}$ and the set V_2 of $v_{k+2}, \ldots, v_{2k+2}$. The set V_1' is composed of $v_1, \ldots, v_{k+1}$ and the set V_2' of $v_{k+2}, \ldots, v_{2k+1}$. Since $\mu'(v_{k+1}) = 0 = \mu(v_{k+1})$, we deduce that $\sum_{v \in V_1} \mu(v) = \sum_{v \in V_2} \mu(v)$ and $\sum_{v \in V_1'} \mu'(v) = \sum_{v \in V_2'} \mu'(v)$. Hence, it follows that for each $i \in \{1, 2\}$,

$$\sum_{v \in V_i'} \mu'(v) - \sum_{v \in V_i} \mu(v) = \left(\sum_{j=0}^{k-1} \frac{1}{k} - \frac{1}{2k-j} \right) - \left(\sum_{j=0}^{k} \frac{1}{k+1} - \frac{1}{2k+2-j} \right)$$
$$= \frac{1}{2k+2} + \frac{1}{2k+1} - \frac{1}{k+1} = \frac{1}{2k+1} - \frac{1}{2k+2} > 0.$$

Now, fix $i \in \{3, \ldots, t-1\}$. Since $T \in \mathcal{F}_{\Delta,k+1}(n)$, there is at most one leaf of T with depth less than $k+1$, which necessarily belongs to V_t. Thus the monotone path P_i induced by V_i in T starts from a leaf of depth $k+1$. Similarly, the monotone path P_i' induced by V_i' in T' starts from a leaf of depth k. Observe that $|V_i| = |V_i'| \in \{1, \ldots, k\}$. Consequently, setting $\ell := |V_i|$, we deduce that

$$\sum_{v \in V_i'} \mu'(v) - \sum_{v \in V_i} \mu(v) = \left(\frac{\ell}{k} - \sum_{j=0}^{\ell-1} \frac{1}{2k-j} \right) - \left(\frac{\ell}{k+1} - \sum_{j=0}^{\ell-1} \frac{1}{2k+2-j} \right)$$
$$= \frac{\ell}{k(k+1)} - \sum_{j=0}^{\ell-1} \frac{1}{2k-j} + \sum_{j=-2}^{\ell-3} \frac{1}{2k-j}$$
$$= \frac{\ell}{k(k+1)} - \frac{1}{2k+1-\ell} - \frac{1}{2k+2-\ell} + \frac{1}{2k+2} + \frac{1}{2k+1}$$
$$= \frac{\ell \cdot f(k,\ell)}{2k(k+1)(2k+1)(2k+1-\ell)(2k+2-\ell)},$$

where $f(k,\ell) := 8k^3 - k^2(12\ell - 20) + k(4\ell^2 - 17\ell + 15) + 2\ell^2 - 6\ell + 4$. As a function of $\ell \in [1,k]$, we see that $f(k,\ell)$ is decreasing so $f(k,\ell) \geqslant f(k,k) = 5k^2 + 9k + 4$, which is positive.

It remains to consider the sets V_t and V_t'. Note that $V_t \subseteq V_t'$. Therefore, the exact same reasoning as above applies, using $|V_t|$ for ℓ and ignoring the nodes in $V_t' \setminus V_t$, which is possible as $\mu'(v) \geqslant 0$ for every node v. Consequently, $E_1(T') > E_1(T)$, which is a contradiction. We conclude that T is a tree of diameter $2k_0$.

Now if $n = \nu(\Delta, k_0)$, then T is the full Δ-regular tree F_{Δ,k_0}, which is the unique element of $\mathscr{T}_{\Delta,k_0}(n)$. Otherwise, $n < \nu(\Delta, k_0)$ and, in particular, the vector $(n_1(T), \ldots, n_{k_0}(T))$ must be an optimal solution to the problem (P) with parameters k_0, $n-1$, Δ and $\alpha_i := \frac{1}{k_0+i}$ for $i \in \{1, \ldots, k_0\}$. Proposition 5.13 thus implies that $(n_1(T), \ldots, n_{k_0}(T))$ is uniquely defined and corresponds to the sizes of the layers of a tree in $\mathcal{F}_{\Delta,k_0}(n)$. We infer that T belongs to $\mathcal{F}_{\Delta,k_0}(n)$, which finishes the proof of Theorem 5.8.

We conclude by pointing out that valid operations provide an efficient algorithmic way of building all possible networks in $\mathscr{T}^*_{n,\Delta}$. We also notice that medians in tree networks with fixed maximum degree Δ need not have degree Δ.

6

CENTRALIZATION OF TRANSMISSION IN NETWORKS

In graph theory, centrality refers to indices which identify the most important (the most central) vertices within a graph. Those most commonly used measures are betweenness centrality, closeness centrality, degree and eccentricity (for more discussion see Section 3.3 on page 21). The *centralization* of a graph is a measure of how central its most central vertex is with respect to how central all the other vertices are (see Section 3.3.2 on page 24 for more discussion).

Transmission of a particular vertex $v \in V(G)$ (in some literature also called *farness* or *vertex-Wiener index*) is defined as a sum of the lengths of all shortest paths between chosen vertex and all other vertices in G (see Section 3.4.5 on page 30). In this chapter we focus to the centralization of transmission.

6.1 BASIC NOTIONS

Let us restate the core definitions, relevant to what follows. The *transmission* $W(v)$ of a vertex $v \in V(G)$ is defined as

$$W(v) = \sum_{u \in V(G)} d_G(u,v).$$

Transmission centralization of a vertex $v \in V(G)$ is obtained by applying Freeman's notion of the centralization to the transmission, formally

$$W_1(v) \quad = \sum_{u \in V(G) \setminus \{v\}} (W(v) - W(u)) = n \cdot W(v) - 2W(G), \tag{6.1}$$

where $W(G)$ is the Wiener index of a graph G. In order to compare centralization values of graphs with different sizes, Freeman in the definition of centralization originally used a normalized formula, dividing expression (6.1) by the theoretically largest such sum of differences in any graph from the given class of graphs [52]. Since throughout this chapter the size of our graph is of constant size, we omit the normalizing denominators.

Several aspects of correlation between Wiener index and betweenness centrality are presented in the paper of Caporossi et al. [29], where authors assign betweenness-related weights to edges of a graph that sum up to its Wiener index. For graphs with fixed order they also find extremal graphs for lower and upper bounds of betweenness centrality. A theorem of Wiener [142], shows how the Wiener index of a tree is decomposed into (easily calculable) edge contributions. In [130], authors introduce a vertex-version of this theorem for general graphs by using the correlation of Wiener index to betweenness centrality.

Among all graphs on n vertices $\mathcal{G}_n$, those that achieve maximum or minimum Wiener centralization value will be called *extremal graphs.* Throughout the chapter we assume that $n > 1$. Instead of $W(v)$ and $W_1(v)$ we will sometimes also write $W(v, G)$ and $W_1(v, G)$, to emphasize the underlying graph we are dealing with. The *eccentricity* of a vertex w is defined as $\max_{v \in V(G)} d_G(w, v)$.

The chapter is structured as follows. In section two, we present the structure of graphs that attain maximal Wiener centralization while in section three we focus on the lower bound. In the concluding chapter we give some ideas for possible future work.

6.2 UPPER BOUND OF TRANSMISSION CENTRALIZATION

In lemmas that follow, we assume that G is a connected graph on n vertices that maximizes transmission centralization among all graphs in $\mathcal{G}_n$. Also, let $w \in V(G)$ be a vertex at which transmission centralization is maximized and let d be the eccentricity of the vertex w. By the choice of w, it is easy to see that for any $t \in V(G)$ we have

$$W(w, G) \quad \geqslant \quad W(t, G) \quad \text{and} \quad W_1(w, G) \geqslant 0. \tag{6.2}$$

Let $L_i := \{v \in G;\, d_G(v, w) = i\}$ be the set of vertices at distance i from w in G, and let $l_i = |L_i|$. We say that L_i is the i-th layer from w. Note that $L_0 = \{w\}$.

Lemma 6.1. *Let i be a non-negative integer. Then vertices in L_i and L_{i+1} induce a complete graph.*

Proof. Assume that there exist two non-connected vertices $u, v \in L_i \cup L_{i+1}$ that violate the claim of this lemma. It is easy to see that adding an edge uv does not affect the value of $W(w)$. On the other hand, introducing the edge uv (or any new edge) always decreases Wiener index of the whole graph. Therefore, introducing the edge uv increases expression (6.1), a contradiction. □

A layer is *trivial* if it is comprised of one vertex.

Lemma 6.2. *Layers $L_1, L_2, \ldots, L_{\lfloor n/2 \rfloor - 1}$ are trivial.*

Proof. Let $s = \lfloor \frac{n}{2} \rfloor$. We proceed with contradiction assuming that some of these layers in G is non-trivial. We prove the claim by introducing an operation that iteratively transform nearest $s-1$ vertices from w into a path, increasing its transmission centralization at each step.

We now describe the operation. Let $i \leqslant s-1$ be the smallest integer such that L_i is a non-trivial layer, and let $v \in L_i$. We construct a new graph G' from G by removing the vertex v and attaching it to the vertex w. Since v is a leaf in G, graph G' clearly remains connected. We prove the claim by showing that $W_1(v, G') > W_1(w, G)$. First notice that

$$W(G') - W(G) = W(v, G') - W(v, G) < (i+1) \cdot (n-i-1).$$

Centralization measures calculate the sum of differences in centrality between the most central vertex in a graph and all other vertices, thus every centrality measure can have its own centralization measure Furthermore, it is important to note that $W(w, G') = W(w, G) - i + 1$ and $W(v, G') = W(w, G') + n - 2$, therefore

$$W(v, G') \quad = \quad W(w, G) + n - i - 1.$$

With help of the above inequalities we now estimate the difference of transmission centralization of optimal vertices of G and G'. Indeed

$$\begin{aligned} W_1(v, G') - W_1(w, G) &= n \cdot \left(W(v, G') - W(w, G)\right) - 2\left(W(G') - W(G)\right) \\ &> n \cdot (n-i-1) - 2 \cdot (n-i-1) \cdot (i+1) \\ &= (n-i-1) \cdot (n-2i-2) \geqslant 0, \end{aligned}$$

where in the final inequality we used the fact that $i \leqslant s-1$. Described operation clearly improves transmission centralization and repeating this process yields the result of the lemma. □

In the next lemma, we show that somewhere "far" from w, there exists $K_{\lceil n/2 \rceil}$ as a subgraph of G.

Lemma 6.3. *If $n > 4$, then the last two layers from w together contain at least $\lceil n/2 \rceil$ vertices.*

Proof. By Lemma 6.2, layers L_0 and L_1 are trivial, implying $d \geqslant 2$. Suppose that the claim is false, i.e. $l_{d-1} + l_d \leqslant \lceil \frac{n}{2} \rceil - 1$, let $v \in L_d$ and let G' be a graph obtained from G by introducing edges from v to all vertices in L_{d-2}; in other words, we move the vertex v from L_d to L_{d-1}. We derive the contradiction by showing that $W_1(w, G') > W_1(w, G)$.

Notice that $W(w, G') - W(w, G) = -1$ and

$$W(G) - W(G') = W(G, v) - W(G', v) = n - l_{d-1} - l_d.$$

Now, from (6.1) it clearly follows

$$\begin{aligned} W_1(w,G') - W_1(w,G) &= n \cdot \big(W(w,G') - W(w,G)\big) + 2\left(W(G) - W(G')\right) \\ &= n - 2(l_{d-1} + l_d) \geqslant n - 2 \cdot \left\lceil \frac{n}{2} \right\rceil + 2 > 0, \end{aligned}$$

a contradiction. □

The result from Lemma 6.3 can be slightly improved if n is odd. We do that in the next lemma.

Lemma 6.4. *If $n \geqslant 5$ is odd, then L_d contains at least $\lceil \frac{n}{2} \rceil - 1$ vertices, and L_{d-1} is trivial.*

Proof. Define s such that $n = 2s + 1$. By Lemma 6.3 we may assume that $\sum_{i=0}^{d-2} l_i \leqslant s$. Suppose that the layer L_{d-1} is not trivial and consider the following operation: Construct a graph G' from G by moving an arbitrary vertex $v \in L_{d-1}$ to L_d (i.e. removing all edges from v to its neighbours in L_{d-2}). Observe that $W(w,G') - W(w,G) = 1$ and note that

$$W(G') - W(G) \quad = \quad W(v,G') - W(v,G) = \sum_{i=0}^{d-2} l_i \leqslant s,$$

therefore by (6.1) and by the fact that n is odd, it holds

$$\begin{aligned} W_1(w,G') - W_1(w,G) &= n \cdot \big(W(w,G') - W(w,G)\big) - 2\left(W(G') - W(G)\right) \\ &\geqslant n - 2 \cdot s > 0. \end{aligned}$$

Clearly, this operation improves transmission centralization of the vertex w, contradicting the choice of G and w. □

The reader might notice that by the above lemmas we are close to describing extremal graphs for transmission centralization. Indeed, almost half of vertices closest to w must form a path. On the other hand, last half of vertices must lie in the last two layers. In what follows we fully characterize the structure of extremal graphs in terms of transmission centralization, but first let us define a notation that will be useful for describing them.

Definition 6.5. For positive integers a, b and $c \leqslant b$, let $\mathrm{PK}(a,b,c)$ be a connected graph on $a + b$ vertices comprised from a path P_a and a clique K_b such that one of the end-vertices of P_a is connected to c vertices of K_b. Thus, $\mathrm{PK}(a,b,c)$ contains $a - 1 + \binom{b}{2} + c$ edges. Few examples are depicted in Figure 6.1.

The following lemma characterizes the structure of extremal graphs for odd n.

Lemma 6.6. *If n is odd, then G is isomorphic to* $\mathrm{PK}\left(\left\lfloor \frac{n}{2} \right\rfloor, \left\lceil \frac{n}{2} \right\rceil, 1\right)$.

Proof. If $n = 3$, the claim is easy to verify. Among two members of $\mathcal{G}_3$, namely K_3 and P_3, centralization is maximized at P_3, which is isomorphic to $\mathrm{PK}(1, 2, 1)$. Suppose again that $n = 2s + 1$ for some integer $s \geqslant 2$. In this case, Lemmas 6.2, 6.3 and 6.4 determine that first s layers are trivial and the last layer contains at least s vertices. So G could be one of the graphs $\mathrm{PK}(s, s+1, 1)$ and $\mathrm{PK}(s, s+1, s+1)$. We now calculate the transmission centralization for each of them separately. In both cases we consider the distances of all pairs by partitioning the vertex set to two sets. The first set is consisted of the layers inducing a path while the remaining two layers are in the second set (inducing a complete graph). First consider the maximum transmission centrality and Wiener index of $\mathrm{PK}(s, s+1, s+1)$

$$\begin{aligned} W(w, \mathrm{PK}(s, s+1, s+1)) &= \binom{s}{2} + s(s+1) = \frac{3}{2}s^2 + \frac{1}{2}s, \text{ and} \\ W(\mathrm{PK}(s, s+1, s+1)) &= W(P_s) + W(K_{s+1}) + (s+1) \cdot \binom{s+1}{2} \\ &= \frac{2s^3}{3} + \frac{3s^2}{2} + \frac{5s}{6}, \end{aligned}$$

from which we derive its transmission centralization

$$\begin{aligned} W_1(w, \mathrm{PK}(s, s+1, s+1)) &= (2s+1) \cdot \left(\frac{3}{2}s^2 + \frac{1}{2}s\right) - 2 \cdot \left(\frac{2}{3}s^3 + \frac{3}{2}s^2 + \frac{5}{6}s\right) \\ &= \frac{5}{3}s^3 - \frac{1}{2}s^2 - \frac{7}{6}s. \end{aligned}$$

Finally, consider the maximum transmission centrality and Wiener index of graph $\mathrm{PK}(s, s+1, 1)$;

$$\begin{aligned} W(w, \mathrm{PK}(s, s+1, 1)) &= \binom{s+1}{2} + (s+1)s = \frac{3}{2}s^2 + \frac{3}{2}s; \\ W(\mathrm{PK}(s, s+1, 1)) &= W(P_{s+1}) + W(K_s) + s \cdot \binom{s+2}{2} \\ &= \frac{2s^3}{3} + \frac{5s^2}{2} + \frac{5s}{6}, \end{aligned}$$

from where we obtain

$$W_1(w, \mathrm{PK}(s, s+1, 1)) = \frac{5}{3}s^3 - \frac{1}{2}s^2 - \frac{1}{6}s > W_1(w, \mathrm{PK}(s+1, s, 1)),$$

which concludes the proof. □

We now proceed with a lemma that characterizes the structure of extremal graphs for even n.

Figure 6.1: The only connected members of $\mathcal{G}_4$ that are consistent with Lemma 6.1.

Lemma 6.7. *If n is even, then G is isomorphic to* $\mathrm{PK}\left(\frac{n}{2}, \frac{n}{2}, i\right)$ *for some* $i \in \left[1, \frac{n}{2}\right]$.

Proof. Again we start the proof by verifying some small non-trivial graphs. For $n = 2$, graph $\mathrm{PK}(1, 1, 1)$ is the only connected graph on 2 vertices. In $\mathcal{G}_4$, there are two connected graphs that are consistent with Lemmas 6.1 and 6.2. These are depicted in Figure 6.1. By easy calculation one can indeed verify that both $\mathrm{PK}(2, 2, 1)$ and $\mathrm{PK}(2, 2, 2)$ attain maximal transmission centralization. Let us assume now that $n = 2s$ for some integer $s \geqslant 3$ and note that graphs $\mathrm{PK}(s, s, i)$ have layers from w of sizes $1, 1, \ldots, 1, i, j$, with $i + j = s$. By Lemmas 6.1–6.3, it is enough to show that all such graphs have the same transmission centralization values. First consider the value of $W(w, \mathrm{PK}(s, s, i))$. By definition,

$$\begin{aligned} W(w, \mathrm{PK}(s, s, i)) &= 1 + 2 + \cdots + (s-1) + s \cdot i + (s+1) \cdot j \\ &= \frac{3}{8}n^2 - \frac{1}{4}n + j. \end{aligned} \tag{6.3}$$

Let us now calculate Wiener index of $\mathrm{PK}(s, s, i)$. Using similar approach as in previous lemma, observe that

$$\begin{aligned} W(\mathrm{PK}(s, s, i)) &= W(P_s) + W(K_s) + i \cdot \frac{s \cdot (s+1)}{2} + j \cdot \frac{(s+1) \cdot (s+2)}{2} - j \\ &= \binom{s+1}{3} + \binom{s}{2} + s \cdot \binom{s+1}{2} + j \cdot s \\ &= \frac{j \cdot n}{2} + \frac{n^3}{12} + \frac{n^2}{4} - \frac{n}{3}. \end{aligned} \tag{6.4}$$

Combining (6.3) and (6.4), it follows that the transmission centralization value of these graphs equals to

$$\begin{aligned} W_1(w, \mathrm{PK}(s, s, i)) &= nW\left(w, U_n^{i,j}\right) - 2W\left(U_n^{i,j}\right) \\ &= \frac{5n^3}{24} - \frac{3n^2}{4} + \frac{2n}{3}, \end{aligned}$$

which is independent on the distribution of i and j. □

By Lemmas 6.1, 6.6 and 6.7, we obtain the main theorem of this chapter.

Parity	G	$W_1(w, G)$
Odd	$\mathrm{PK}\left(\lfloor\frac{n}{2}\rfloor, \lceil\frac{n}{2}\rceil, 1\right)$	$\frac{5}{24}n^3 - \frac{3}{4}n^2 + \frac{19}{24}n - \frac{1}{4}$
Even	$\mathrm{PK}\left(\frac{n}{2}, \frac{n}{2}, i\right)$	$\frac{5}{24}n^3 - \frac{3}{4}n^2 + \frac{2}{3}n$

Table 6.1: Values of transmission centralization for extremal graphs on odd and even number of vertices.

Theorem 6.8. *In the family of $\mathcal{G}_n$, graph that maximizes transmission centralization is isomorphic to*

- $\mathrm{PK}\left(\frac{n}{2}, \frac{n}{2}, i\right)$ *for some* $i \in \left[1, \frac{n}{2}\right]$ *for even* n*, and*
- $\mathrm{PK}\left(\lfloor\frac{n}{2}\rfloor, \lceil\frac{n}{2}\rceil, 1\right)$ *for odd* n*.*

The summary of optimal transmission centralization values is shown in Table 6.1.

6.3 LOWER BOUND OF TRANSMISSION CENTRALIZATION

In this section, we characterize the graphs in $\mathcal{G}_n$ that attain minimum value of transmission centralization. Henceforth we assume that G achieves minimum transmission centralization among all graphs in $\mathcal{G}_n$. Also, let $w \in V(G)$ be a vertex at which transmission centralization is minimized. Let us first start with a simple lemma, which shows that G is acyclic.

Lemma 6.9. *Graph G is a tree.*

Proof. Suppose that G is not a tree and let T be a breadth-first search tree of G rooted at w. Note that removing an edge from a simple graph increases its Wiener index, therefore $W(T) > W(G)$. By the fact that $W(w, T) = W(w, G)$, the claim follows from

$$\begin{aligned} & W_1(w, T) - W_1(w, G) \\ = \; & n \cdot (W(w, T) - W(w, G)) - 2 \cdot (W(T) - W(G)) \\ = \; & -2 \cdot (W(T) - W(G)) < 0. \end{aligned}$$

□

To proceed with the proof of the lower bound, we will need some additional notations. From now on we will consider G as a tree, rooted at vertex w. For any $x \in V(G)$, let T_x be the subtree of G rooted at x, and let $t_x := |V(T_x)|$. We continue with a lemma stating that G is a subdivision of a star.

Lemma 6.10. *Every vertex of G, distinct from w, has degree less or equal to* 2.

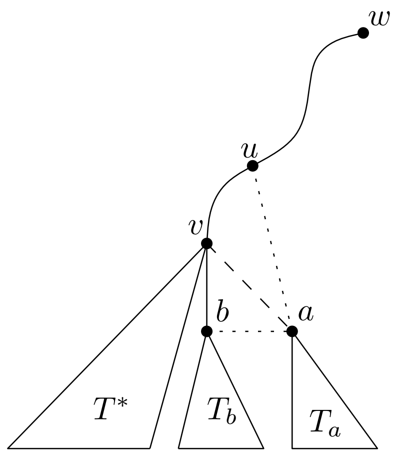


Figure 6.2: Notations around a vertex v. Dashed edge is removed in G', and one of the dotted edges is added.

Proof. Suppose that there exists a vertex v that violates the claim, and let u be its parent node. Since $\deg(v) \geqslant 3$, there exist at least two additional neighbors of v, namely a and b. Let T^* be a subtree in G, induced by vertices $V(T_v) \setminus (V(T_a) \cup V(T_b))$ and let $t^* = |V(T^*)|$. Note that T^* is possibly comprised only of v. For better illustration, see Figure 6.2. Note that parameters t_a, t_b and t^* are strictly positive and, depending on those values, we derive a contradiction by reattaching T_a to b or u.

Suppose first that $n > 4t_b + 2t_a$. In this case, let G' be a graph obtained from G by disconnecting a from v and connecting it to the vertex b. We will now compare Wiener indices of both graphs and conclude by showing that $W_1(w, G) > W_1(w, G')$. First notice that

$$d_{G'}(w,z) - d_G(w,z) = \begin{cases} 1 & \text{if } z \in T_a \\ 0 & \text{otherwise,} \end{cases}$$

therefore it is clear that $W(w, G') - W(w, G) = t_a$. By similar argument it is easy to see that the distance between any two vertices changes only when precisely one of them is a member of T_a. We can therefore conclude

$$\begin{aligned} W(G') - W(G) &= \sum_{c \in V(T_a)} \left[\left(\sum_{d \in \mathcal{A}} 1 \right) + \left(\sum_{d \in V(T_b)} -1 \right) \right] \\ &= t_a \cdot (n - t_a - 2t_b)\,, \end{aligned}$$

where $\mathcal{A} = V(G) \setminus (V(T_a) \cup V(T_b))$. We now calculate the change of the transmission centralization of the vertex w

$$\begin{aligned} W_1(w, G') - W_1(w, G) &= n \cdot t_a - 2 \cdot t_a (n - t_a - 2t_b) \\ &= t_a \cdot (2t_a + 4t_b - n), \end{aligned}$$

which by our assumption implies the claim for this case.

Consider now that $n \leqslant 4t_b + 2t_a$. In this case, let G' be a graph obtained from G by disconnecting a from v and connecting it to the vertex u. We will again compare both graphs and conclude by showing that $W_1(w, G) > W_1(w, G')$. First notice that

$$d_{G'}(w,z) - d_G(w,z) = \begin{cases} -1 & \text{if } z \in T_a \\ 0 & \text{otherwise,} \end{cases}$$

and hence $W(w, G') - W(w, G) = -t_a$. Again, the distance between any two vertices changes only when one of them is a member of T_a. We can therefore conclude

$$\begin{aligned} W(G') - W(G) &= \sum_{c \in V(T_a)} \left[\left(\sum_{d \in \mathcal{A}} 1 \right) + \left(\sum_{d \in \mathcal{B}} -1 \right) \right] \\ &= t_a \cdot (t^* + t_b) - t_a \cdot (n - t_a - t_b - t^*) \\ &= t_a \cdot (2t^* + 2t_b - n + t_a), \end{aligned}$$

where $\mathcal{A} = V(T^*) \cup V(T_b)$ and $\mathcal{B} = V(G) \setminus (V(T^*) \cup V(T_b) \cup V(T_a))$. From these facts, we can now calculate the change of the transmission centralization at the vertex w

$$W_1(w, G') - W_1(w, G) = t_a \cdot (n - 4t^* - 4t_b - 2t_a) < t_a \cdot (n - 4t_b - 2t_a)$$

which concludes the proof of the claim. □

By the above lemma, G is a subdivision of a k-star, and therefore we introduce the following definition. For some non-decreasing sequence of positive integers $\alpha_1, \alpha_2, \ldots, \alpha_k$ let $P(\alpha_1, \alpha_2, \ldots, \alpha_k)$ be a tree on $1 + \sum_{i=1}^{k} \alpha_i$ vertices comprised from k paths $P_{\alpha_1}, P_{\alpha_2}, \ldots, P_{\alpha_k}$ such that one of the end-vertices of each path is connected to the additional vertex of degree k. We label this vertex with w. The next claim will help us calculate its transmission centralization.

Lemma 6.11. *The transmission centralization of the graph $P(\alpha_1, \alpha_2, \ldots, \alpha_k)$ equals*

$$\frac{1}{6} \cdot \sum_{i=1}^{k} \left[4\alpha_i^3 + (6 - 3n)\, \alpha_i^2 + (2 - 3n)\, \alpha_i \right]. \tag{6.5}$$

Proof. It is easy to sum all distances from w to all other vertices. Indeed,

$$W\left(w, P(\alpha_1, \alpha_2, \ldots, \alpha_k)\right) = \sum_{i=1}^{k} \binom{\alpha_i + 1}{2}.$$

The Wiener index of $P(\alpha_1, \alpha_2, \ldots, \alpha_k)$ is comprised of two types

- distances where both ends are from the same path P_{α_i}, these sum up to Wiener index of a path of length α_i,
- distances between $u \in P_{\alpha_i}$ and $v \notin P_{\alpha_i}$ for some i; to avoid multiple counting here we count at i only contributions made on the path P_{α_i}.

Hence, the Wiener index of the graph $P(\alpha_1, \alpha_2, \ldots, \alpha_k)$ equals

$$\begin{aligned} W\left(P(\alpha_1, \alpha_2, \ldots, \alpha_k)\right) &= \sum_{i=1}^{k} \left[W\left(P_{\alpha_i}\right) + \sum_{v \in P_{\alpha_i}} \sum_{v' \notin P_{\alpha_i}} d(w, v) \right] \\ &= \sum_{i=1}^{k} \left[\binom{\alpha_i + 1}{3} + (n - \alpha_i) \cdot \binom{\alpha_i + 1}{2} \right]. \end{aligned}$$

Plugging both expressions to (6.1) leads to the result of the claim. □

What remains to be done is to find k and appropriate values of $\alpha_1, \ldots, \alpha_k$, such that the expression (6.5) is minimized.

Let $G = P(\alpha_1, \ldots, \alpha_k)$ be a graph on at least six vertices and define graph G' from G by changing $\alpha_i \to \alpha_i - 1$ and $\alpha_j \to \alpha_j + 1$. Using Lemma 6.11 it is easy to conclude that the difference of transmission centralization between these two graphs is

$$W_1(w, G') - W_1(w, G) = (\alpha_i - \alpha_j - 1)(n - 2\alpha_i - 2\alpha_j - 2). \tag{6.6}$$

Note that the expression also holds if $\alpha_j = 0$, which corresponds to introducing a new branch of length one by removing one vertex from a branch of length α_i. Let us derive some properties of G by use of expression (6.6).

Claim 1. *If $n \geqslant 4$, then $\alpha_k \leqslant \frac{n-2}{2}$.*

Proof. Observe that in case $\alpha_k = 1$, the claim is trivially satisfied, so we assume that $\alpha_k \geqslant 2$. Suppose that $\alpha_k > \frac{n-2}{2}$. Now, change $\alpha_k \to \alpha_k - 1$ and introduce a new branch to w by attaching a new leaf to w. From (6.6) it clearly follows that

$$W_1(w, G') - W_1(w, G) = (\alpha_k - 1)(n - 2\alpha_k - 2),$$

with $i = k$ and $\alpha_j = 0$. By our assumption, the resulting graph has smaller transmission centralization value, which contradicts the choice of G. □

Thus, by the above claim it follows that $n \geqslant 4$ implies $k \geqslant 3$.

Claim 2. $\alpha_1 + \alpha_2 \geqslant \frac{n-2}{2}$.

Proof. Indeed, if $\alpha_1 + \alpha_2 < \frac{n-2}{2}$, replace α_1 and α_2 by $\alpha_1 - 1$ and $\alpha_2 + 1$, respectively. By our assumption, $n - 2\alpha_1 - 2\alpha_2 - 2$ is positive, whereas $\alpha_1 - \alpha_2 - 1$ is negative. Hence, (6.6) is negative and the obtained graph contradicts the choice of G. □

It follows that $\lceil \frac{n-2}{4} \rceil \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_k$, and that $n \geqslant 7$ implies $3 \leqslant k \leqslant 4$. Indeed, if $k \geqslant 5$ then

$$n - 1 \geqslant \sum_{i=1}^{5} \alpha_i \geqslant \left\lceil \frac{5n-10}{4} \right\rceil ,$$

which is a contraction whenever $n > 6$. Finally, we have the third property which shows that all branches are of approximately the same length.

Claim 3. $\alpha_k - \alpha_1 \leqslant 1$.

Proof. Suppose otherwise and observe the same operation as in the expression (6.6) with $\alpha_1 \to \alpha_1 + 1$ and $\alpha_k \to \alpha_k - 1$. By Claim 2 we have $\alpha_1 + \alpha_k \geqslant \alpha_1 + \alpha_2 \geqslant \frac{n-2}{2}$, therefore $n - 2\alpha_k - 2\alpha_1 - 2$ is strictly negative. Furthermore, by our assumption $\alpha_k - \alpha_1 - 1$ is positive. Hence, (6.6) is negative, which is a contradiction with the choice of G. [1]

□

It remains to decide for each value of n, whether $k = 3$ or $k = 4$. We deal with most of the cases in the following claim.

Claim 4. *If $n \geqslant 9$, then $k = 3$.*

Proof. For $i \in \{3, 4\}$, let G_i be a graph with $k = i$ that minimizes transmission centralization in its minimizing vertex w. For easier notation, let $f(x) = 4x^3 + (6 - 3n)\,x^2 + (2 - 3n)\,x$. Observe that for $W_1(w, G_4) = \frac{1}{6} \cdot (f(\alpha_1) + f(\alpha_2) + f(\alpha_3) + f(\alpha_4))$, depending on the value of $n \pmod 4$, we have four possibilities for $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. Similarly, for $W_1(w, G_3) = \frac{1}{6} \cdot (f(\alpha_1) + f(\alpha_2) + f(\alpha_3))$, we

1 The authors would like to thank prof. Brendan McKay for providing a useful response [102] that simplified the proof of Claim 3.

k	$n \pmod k$	$6 \cdot W_1(w, G_k)$
3	0	$f\left(\frac{n-3}{3}\right) + 2 \cdot f\left(\frac{n}{3}\right) = -\frac{5}{9}n^3 - \frac{1}{3}n^2 + 2n$
	1	$3 \cdot f\left(\frac{n-1}{3}\right) = -\frac{5}{9}n^3 - \frac{1}{3}n^2 + \frac{4}{3}n - \frac{4}{9}$
	2	$2 \cdot f\left(\frac{n-2}{3}\right) + f\left(\frac{n+1}{3}\right) = -\frac{5}{9}n^3 - \frac{1}{3}n^2 + 2n + \frac{16}{9}$
4	0	$f\left(\frac{n-4}{4}\right) + 3 \cdot f\left(\frac{n}{4}\right) = -\frac{1}{2}n^3 - \frac{3}{4}n^2 + 2n$
	1	$4 \cdot f\left(\frac{n-1}{4}\right) = -\frac{1}{2}n^3 - \frac{3}{4}n^2 + 2n - \frac{3}{4}$
	2	$2 \cdot f\left(\frac{n-3}{4}\right) + 2 \cdot f\left(\frac{n+1}{4}\right) = -\frac{1}{2}n^3 - \frac{3}{4}n^2 + 2n + \frac{9}{4}$
	3	$3 \cdot f\left(\frac{n-2}{4}\right) + f\left(\frac{n+2}{4}\right) = -\frac{1}{2}n^3 - \frac{3}{4}n^2 + 2n + 3$

Table 6.2: All possibilities for $W_1(w, G_4)$ and $W_1(w, G_3)$ depending on the value of $n \pmod 4$ and the value of $n \pmod 3$.

have three possibilities for $\alpha_1, \alpha_2, \alpha_3$. All combinations are listed in Table 6.2. Observing Table 6.2, it is clear that

$$W_1(w, G_4) \geqslant -\frac{1}{12}n^3 - \frac{1}{8}n^2 + \frac{1}{3}n - \frac{1}{8}$$

and

$$W_1(w, G_3) \leqslant -\frac{5}{54}n^3 - \frac{1}{18}n^2 + \frac{1}{3}n + \frac{8}{27}.$$

Setting $n \geqslant 9$, the claim immediately follows. □

From the statements above follows the main theorem of this section.

Theorem 6.12. *Let $n \geqslant 9$ and let $\alpha_1 \leqslant \alpha_2 \leqslant \alpha_3$ be positive integers, such that $\alpha_1 + \alpha_2 + \alpha_3 = n - 1$ and $\alpha_3 - \alpha_1 \leqslant 1$. Than, in the family of $\mathcal{G}_n$, graph that minimizes transmission centralization is isomorphic to $P(\alpha_1, \alpha_2, \alpha_3)$.*

By Lemma 6.10 and Claim 3, we can easily determine the rest of extremal graphs (on less than 9 nodes) that minimize transmission centralization by hand. All these minimizing graphs are listed on Figure 6.3. Let us note that for cases $n \in \{6, 8\}$, the minimizing tree is not unique.

We conclude with the following corollary.

Corollary 6.13. *Let $H \in \mathcal{G}_n$ and $v \in V(H)$. Then*

$$-\frac{5}{54}n^3 + O\left(n^2\right) \leqslant W_1(v) \leqslant \frac{5}{24}n^3 + O\left(n^2\right).$$

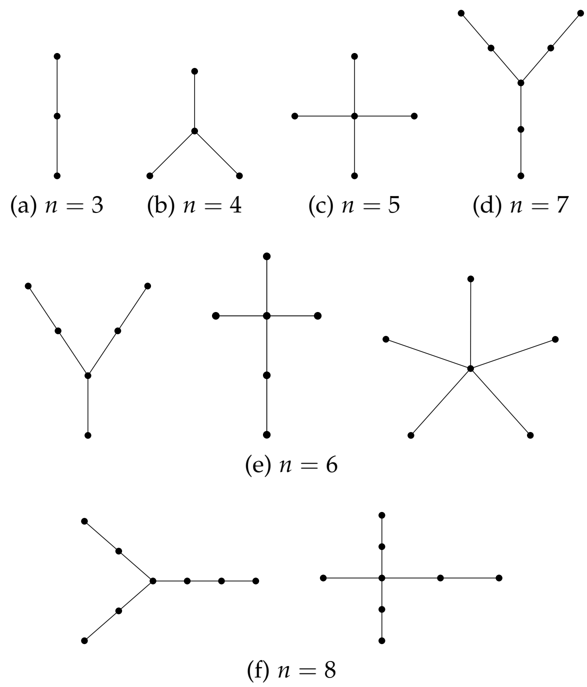


Figure 6.3: The list of extremal graphs with $3 \leqslant n \leqslant 8$ that minimize transmission centralization.

6.4 CONCLUDING REMARKS AND FUTURE WORK

Finding extremal graphs with respect to the centralization measures was previously studied by Butts [28], Everett et al. [48] and Freeman [52]. For most common centrality indices (degree, betweenness, closeness, eccentricity), a graph with maximal centralization is a star. In sections 2 and 3, we presented extremal graphs for transmission centralization, and to our surprise the result was a mixture between a clique and a path or a collection of three paths of almost the same length, glued together in one end-vertex.

It would be interesting to study further

- the extremal graphs for transmission centralization on some other relevant classes of graphs such as trees or graphs with bounded-degree,
- the centralization for other chemical indices.

7

GROUP CENTRALIZATION OF NETWORK INDICES

For many decades in social science research, social networks have been the subject of study. A social network is typically represented as a graph, where individual persons or nodes are represented as vertices, and the relationships between pairs of individuals as edges. *Centrality* is an important concept in studying social networks that provides us with an information how central is the position of an individual (or a small group) within a network. Various vertex-based measures of the centrality have been proposed to determine the relative importance of a vertex within the graph. Among most used centrality indices in network analysis are: degree centrality, betweenness centrality, closeness centrality, eccentricity centrality, Google PageRank, eigenvector centrality and others. In his study, Freeman [52] realized that despite all defined vertex-centrality indices, there was a need for *graph centrality* measure based on differences in point centrality. He defined a centralization index that can be used in combination with any vertex-centrality to determine to what extent some vertex in network stands out from others in terms of given centrality index. For more discussion on networks, centrality or Freeman centralization see Chapter 3 on page 13, Section 3.3 on page 21 or Section 3.3.2 on page 24, respectively.

In this chapter we determine some graphs that maximize group centralization with respect to eccentricity, degree and betweenness centrality measures. Let us first present some historical discussion surrounding our work.

7.1 BACKGROUND

To find an extremal graph G and maximizing subset $S \subset V(G)$ inside from algorithmic point of view can be a time consuming problem. In 2001 Brandes [23] improved the algorithm for calculating betweenness centrality to $O(|V(G)| \cdot |E(G)|)$. Later in 2008 [24], he extended his algorithm to group betweenness and other similar centralities. There are also some efficient heuristics and greedy approaches that can find vertices or groups that are sub-optimal in terms of various centrality measures, see Puzis et al. [121], Dolev et al. [40].

In 1999, Everett and Borgatti [46] introduced the concept of *group centrality* which enables researchers to answer questions such as "how central is the engineering department in the informal influence network of this company?" or "among middle managers in a given organization, which are more central, the men or the women?" With these measures we can also solve the inverse problem: given the network of ties among organization members, how can we form a team that is maximally central? In [46], the authors introduced group centrality for measures of degree, closeness and betweenness centrality, which we use in the thesis. In 2006, Borgatti introduced important group centrality measure (usually called *KPP*) that is motivated by *key players problem* (see [20]). In his paper he focused on finding a set of vertices for the purpose of optimally diffusing something through the network by using selected vertices as seeds, or for maximally fragmenting the network by removing the key nodes. Interestingly, Borgatti claims that previously mentioned group closeness and betweenness are not proper tools to define KPP centrality. He therefore used tools like graph fragmentation and information entropy to define KPP centrality.

Several more concepts of vertex centrality with respect to some subset of vertices have been introduced throughout last decade. In 2003, Smith and White [141] introduced a measure called *personalization* that shows, how central an individual is according to given subset R (group of important people) in given social network. In 2005, *subgraph centrality* has been introduced by Estrada and Rodríguez-Velázquez [45], and characterizes the participation of each node in all subgraphs in a network, which is calculated from the spectra of the adjacency matrix of the network. In the same year Everett and Borgatti [47] introduced another measure (i.e. *core centrality*), where they evaluate the extent to which a network revolves around a core group of nodes. Finally, very recently Bell [17] introduced the concept called *subgroup centrality*, where centrality (of one vertex) is calculated only on restricted set of vertices. Let us remark that all four mentioned centralities in principle measure importance of an individual vertex (with respect to some conditions) and are different from group centrality, proposed in [46].

Knowing all those group centrality measures it is natural to ask how much some choice of central group stands out from all other groups of the same cardinality (with respect to given group centrality index). Following Freeman's approach, we define group centralization notion in Definition 7.1 on the facing page and discuss it further in later chapters.

In the sequel, we will use the following notion. Denote by $\mathcal{G}_n$ the family of non-isomorphic connected graphs on n vertices. Notice that when we consider a graph G, we usually assume $G \in \mathcal{G}_n$. A *star graph* S_n is a tree on $n+1$ vertices, with one vertex of degree n and n leaves. We will use $N(v)$ as a set of vertices in the neighborhood of v. As we deal with group centralization, by $C \subseteq V(G)$ we always denote the group we consider, and in addition we assume c is the size of C, i.e. $c = |C|$. Since $C = V(G)$ always trivially produces zero centrality (and therefore centralization), we will always assume $c < n$. At last, the *distance* from a vertex $x \in V(G)$ to a set of vertices $C \subseteq V(G)$ is defined by $d(x, C) = \min_{x \in C} \{d(x, c)\}$.

The chapter is structured as follows. In Section 7.2 we introduce group centralization notion for arbitrary centrality index, and briefly describe its origin. In Section 7.3, we consider degree group centralization, and characterize extremal-pairs for graph family $\mathcal{G}_n$. In Section 7.4 we deal with eccentricity group centralization in the same graph family for groups of size 2 and describe the corresponding extremal graphs. In Section 7.5, we then do similar for betweenness group centralization. We conclude with posing few open problems in Section 7.6.

7.2 GROUP CENTRALIZATION

In many real life networks, it is intuitively clear that some nodes are more important than others. Also some graphs are more depending on the most central vertices than others. While centrality measures compare the importance of a node within graph, the associated notion of *centralization*, as introduced by Freeman [52] allows us to compare the relative importance of nodes within their respective graphs. He proposed a very general approach with which the centralization of a graph G can be calculated. A clique where every vertex is connected to every other vertex is clearly not very centralized; on the other hand, the star topology, in which only one vertex v is connected to all others and all other vertices are only connected to v is a centralized graph. Thus, one would expect a star to have greater centralization than clique. In a network G, given a centrality index $X : V(G) \to \mathbb{R}$, the *centralization* of a node v is given by

$$X_1(G, v) = \sum_{u \in V(G)} (X(v) - X(u)) . \tag{7.1}$$

Following Freeman's idea, group centralization can be naturally generalized as a measure of how central its most central set of size c is in relation to how central all the other sets of the same cardinality are. Now we state this formally.

Definition 7.1 (Group centralization). Let G be a graph, $C \subseteq V(G)$, $c := |C|$ and let X be a given group centrality measure. The *group centralization* is defined as

$$\mathrm{GX}_1(G, C) = \sum_{S \in \binom{V(G)}{c}} (X(C) - X(S)) . \tag{7.2}$$

Let c, n be fixed integers with $c < n$. Given centrality measure X, a pair (G, C) is called an *extremal-pair*, if C maximizes the group centralization in G among all possible choices of G and C with $|V(G)| = n$ and $|C| = c$, in other words, if

$$\mathrm{GX}_1(G, C) = \max_{G' \in \mathcal{G}_n} \max_{C' \in \binom{V(G')}{c}} \mathrm{GX}_1(G', C').$$

In the definition of centralization, Freeman used a normalized formula, dividing expression (7.1) by the maximum possible sum of differences in point centrality for a graph of n points, resulting

$$\overline{\mathrm{X}}_1(G, v) = \frac{\mathrm{X}_1(G, v)}{\max_{G' \in \mathcal{G}_n} \max_{v' \in V(G')} \mathrm{X}_1(G', v')}. \tag{7.3}$$

In the same way, we can normalize expression (7.2). For an extremal-pair (G^*, C^*), we have

$$\overline{\mathrm{GX}}_1(G, C) = \frac{\mathrm{GX}_1(G, C)}{\mathrm{GX}_1(G^*, C^*)}. \tag{7.4}$$

This is useful for comparing centralization scores of networks with different group sizes. Since in this chapter, we only work with constant group size, we omit the normalizing denominators.

7.3 GROUP DEGREE CENTRALIZATION

For a graph G and a subset of vertices $C \subseteq V(G)$, Everett and Borgatti [46] introduced *group degree* (i.e. *group degree centrality*) as follows:

$$\mathrm{GD}(C) = \left| \bigcup_{v \in C} N(v) \setminus C \right|.$$

For a given graph G, finding a set C of given cardinality k that maximizes $GD(C)$ is $\mathcal{NP}$-hard (see Miyano and Ono [107]). In the next theorem we characterize extremal-pairs of group degree centralization.

Theorem 7.2. *Let (G, C) be an extremal-pair for group degree centralization in $\mathcal{G}_n$. Then, (S_{n-1}, C) is an extremal-pair for group degree centralization in $\mathcal{G}_n$ (for some appropriate $C \subseteq V(S_{n-1})$).*

Proof. For graphs on two vertices, the claim trivially holds, therefore we assume $n \geqslant 3$. For each vertex $v \in V(G)$, define its contribution $g_k(v)$ to be the number of k-sets that dominates v, i.e.

$$\begin{aligned} g_k(v) &= \left| \left\{ X \in \binom{V(G) - v}{k}; X \cap N(v) \neq \emptyset \right\} \right| \\ &= \binom{n-1}{k} - \binom{n - \deg(v) - 1}{k}. \end{aligned}$$

It follows that

$$\sum_{C'\in\binom{V(G)}{k}} \mathrm{GD}(C',G) = \sum_{v\in V(G)} g_k(v) = n\cdot\binom{n-1}{k} - \sum_{v\in V(G)}\binom{n-\deg(v)-1}{k},$$

which implies

$$\mathrm{GD}_1(C,G) = \binom{n}{k}\cdot \mathrm{GD}(C,G) + \sum_{v\in V(G)}\binom{n-\deg(v)-1}{k} - n\cdot\binom{n-1}{k}. \quad (7.5)$$

From (7.5) one can observe that the value of $\mathrm{GD}_1(C,G)$ depends only on $\mathrm{GD}(C,G)$ and on the degree distribution of G. To maximize (7.5), we first maximize $\sum_{v\in V(G)}\binom{n-\deg(v)-1}{k}$ independently, and then show that the maximizing graph also maximize the value of $\mathrm{GD}(C,G)$ (with properly chosen set C).

It is easy to see that $\sum_{v\in V(G)}\binom{n-\deg(v)-1}{k}$ is maximized when the total number of edges in G is minimized. Since G must remain connected, we conclude that $\sum_{v\in V(G)}\deg(v) = 2n-2$, i.e. G is a tree. Furthermore, since $\binom{n-x-1}{k}$ is a convex decreasing function on $x\in[1,n-1]$, the mentioned sum is maximized if and only if there exists $v\in V(G)$ such that $\deg_G(v) = n-1$ and $\deg_G(u) = 1$ whenever $u\in V(G)\setminus\{v\}$. This happens if and only if G is isomorphic to a star S_{n-1}.

It remains to observe that $\mathrm{GD}(C,G)\leqslant n-k$, where equality is attained if and only if C is a dominating set in G. Note that in S_{n-1} any set C containing the star center dominates the whole graph, hence the the maximum of $\mathrm{GD}(C,G)$ can always be attained in the star on n vertices. □

7.4 Group eccentricity

Given a graph G, the eccentricity of a vertex $v\in V(G)$ is the maximum distance from v to any other vertex. In social networks, Hage and Harary [63] introduced the inverse of eccentricity as a centrality measure. Based on this, one can naturally define the *group eccentricity* of a set $C\subset V(G)$ as maximum distance from any vertex from $V(G)\setminus C$ to closest member of C. Formally, for any set $C\subset V(G)$, let

$$\mathrm{GE}(G,C) = \frac{1}{\max_{x\in V(G)\setminus C} d(x,C)}.$$

Remark 7.3. Note that $\mathrm{GE}(G,C)=1$ if and only if C is a dominating set of G.

Proposition 7.4. *There exists an extremal-pair (T,C) for eccentricity such that T is a tree.*

Proof. Suppose that (G, C) is an extremal pair and build a tree T from the graph G according to the following procedure. Assume all edges of G are initially colored black.

1. First introduce a new vertex r in G, and connect it to every vertex of C.
2. Next color all edges of a breadth-first search tree of $G + r$ rooted at r with blue color.
3. Remove r and observe that the remaining blue graph is in fact a forest F with c components.
4. Color $c - 1$ black edges into blue in order to form a tree T, induced on blue edges.

Clearly all distances $d_G(C, v) = d_T(C, v)$ for $v \in V(G) \setminus C$, therefore group eccentricity for C remains the same, on the other hand, for all other $C' \in \binom{V(G)}{|C|} \setminus C$, group eccentricity may decrease or remain the same. Therefore $\mathrm{GE}_1(C, G) \leqslant \mathrm{GE}_1(C, T)$, implying that (T, C) is also an extremal pair, which concludes our argument. □

Let $D(n, l)$ be the *Dandelion* graph on n vertices, consisted of a star S_{n-l} and a path P_l, on vertices $p_0, p_1, \ldots, p_{l-1}$, where p_0 is identified with a star center. An example of $D(17, 8)$ is shown in Fig. 7.1. In order to use the Dandelion graph family as lower bound on group eccentricity centralization, now we give a lower bound of $\mathrm{GE}_1\left(D(n, l), C\right)$, for a chosen set C.

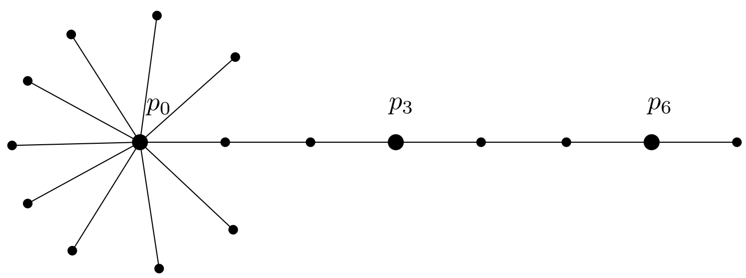


Figure 7.1: A Dandelion graph $D(17, 8)$ with vertices of C enlarged.

Proposition 7.5. *Let $l = 3c - 1$ and $n > l$. Then, it holds*

$$\mathrm{GE}_1\left(D(n, l), C\right) \geqslant \frac{\binom{n}{c} - 1}{2},$$

where $C = \{p_0, p_3, \ldots, p_{l-1}\}$.

Proof. For the sake of simplicity, let $D := D(n,l)$. Since every vertex of D is adjacent or belongs to C, we have that $\mathrm{GE}(D,C) = 1$. From other side, for any other c-set C' distinct from C, there is always a vertex of D on distance $\geqslant 2$ from C'. This implies that $\mathrm{GE}(D,C') \leqslant \frac{1}{2}$. Hence

$$\mathrm{GE}_1(D,C) = \sum_{C' \in \binom{V(D)}{c}} (\mathrm{GE}(D,C) - \mathrm{GE}(D,C')) \geqslant \frac{\binom{n}{c} - 1}{2}.$$

□

From this result, the following corollary clearly holds.

Corollary 7.6. *Let G be a connected graph on n vertices, and let $n > 3c - 1$. For any extremal-pair (G,C), it holds $\mathrm{GE}(G,C) = 1$, i.e. C dominates G.*

From now on, we will focus on extremal-pairs in the family of trees on n vertices $\mathcal{T}_n$, with $c = 2$, where we denote $C = \{x,y\}$. We analyze the distance between x and y in the following lemma.

Lemma 7.7. *Let (G,C) be an extremal-pair with $C = \{x,y\}$ and $n \geqslant 6$. Then $d_G(x,y) = 3$.*

Proof. It is easy to see that $d \leqslant 3$, otherwise on the (x,y)-path there would exist a vertex v such that $d(v,C) > 1$, which is contradiction to Corollary 7.6. Now we prove that $d \geqslant 3$ by considering all three remaining cases, depending on $d(x,y)$, see Fig. 7.2. In all three cases, the graph from Fig. 7.2 is fully defined by the number of leaves attached at x and y, denoted with a and b, respectively, as shown in Fig. 7.2. Without loss of generality we will assume that $a \geqslant b$.

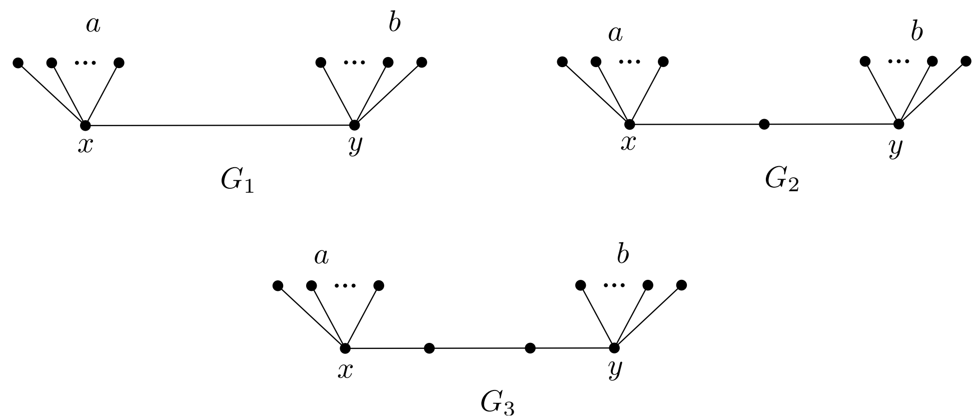


Figure 7.2: The three possible cases regarding $d(y,x)$.

Case $d(x,y) = 1$: Since $a + b \geqslant 4$, observe the value of $\mathrm{GE}_1(G_a, C)$ in the following table.

CONDITION:	VALUE OF $\mathrm{GE}_1(G_a, C)$:
if $a \geqslant 2$ and $b \geqslant 2$:	$\left(\binom{a}{2} + \binom{b}{2}\right) \cdot \frac{2}{3} + (ab + 2a + 2b) \cdot \frac{1}{2}$
if $a \geqslant 3$ and $b = 1$:	$\binom{a}{2} \cdot \frac{2}{3} + (3a + 1) \cdot \frac{1}{2}$
if $a \geqslant 4$ and $b = 0$:	$\frac{1}{2} \cdot \binom{a+1}{2}$

Since the case $a \geqslant 2, b \geqslant 2$ always results in bigger group eccentricity, we can assume

$$\mathrm{GE}_1(G_a, C) = \left(\binom{a}{2} + \binom{b}{2}\right) \cdot \frac{2}{3} + (ab + 2a + 2b) \cdot \frac{1}{2}.$$

Since $a + b = n - 2$ is fixed, after some calculations, the expression can be simplified to

$$\mathrm{GE}_1(G_a, C) = \frac{n^2}{3} - \frac{2n}{3} - \frac{ab}{6}.$$

Clearly, the expression above is maximized, when $a = n - 4$ and $b = 2$, so, we have

$$\mathrm{GE}_1(G_a, C) = \frac{1}{3}n^2 - n + \frac{4}{3}.$$

Case $d(x, y) = 2$: Observe the value of $\mathrm{GE}_1(G_b, C)$ in the following table.

CONDITION:	VALUE OF $\mathrm{GE}_1(G_b, C)$:
if $a \geqslant 2$ and $b \geqslant 2$:	$\left(\binom{a}{2} + \binom{b}{2}\right) \cdot \frac{3}{4} + (a + b) \cdot \frac{2}{3}$ $+ (ab + a + b + n - 1) \cdot \frac{1}{2}$
if $a \geqslant 2$ and $b = 1$:	$\binom{a}{2} \cdot \frac{3}{4} + (a + 1) \cdot \frac{2}{3} + (2a + n - 1) \cdot \frac{1}{2}$
if $a \geqslant 3$ and $b = 0$:	$\binom{a}{2} \cdot \frac{2}{3} + (3a + 1) \cdot \frac{1}{2}$

Again, the case $a \geqslant 2, b \geqslant 2$ always results in bigger group eccentricity and we assume

$$\begin{aligned} \mathrm{GE}_1(G_b, C) \quad &= \quad \left(\binom{a}{2} + \binom{b}{2}\right) \cdot \frac{3}{4} + (a + b) \cdot \frac{2}{3} \\ &\quad + (ab + a + b + n - 1) \cdot \frac{1}{2}. \end{aligned}$$

Since $a+b=n-3$, after some calculations, the expression can be simplified to

$$\begin{aligned}\mathrm{GE}_1(G_b,C) &= \frac{3}{8}\left(a^2+\frac{4}{3}ab+b^2\right)+\frac{31}{24}n-\frac{23}{8}\\ &= -\frac{ab}{4}+\frac{3}{8}n^2-\frac{9}{4}n+\frac{27}{8}+\frac{31}{24}n-\frac{23}{8}\\ &= -\frac{ab}{4}+\frac{3}{8}n^2-\frac{23}{24}n+\frac{1}{2}.\end{aligned}$$

Maximizing the expression, it is clear that $a=n-5$ and $b=2$, so we have

$$\mathrm{GE}_1(G_b,C)=\frac{3}{8}n^2-\frac{35}{24}n+3.$$

Case $d(x,y)=3$**:** Observe the value of $\mathrm{GE}_1\,(G_c,C)$ in the following table.

CONDITION:	VALUE OF $\mathrm{GE}_1\,(G_c,C)$:
if $b \geqslant 1$:	$\left(\binom{a}{2}+\binom{b}{2}\right)\cdot\frac{4}{5}+(a+b)\cdot\frac{3}{4}$ $+(a+b+2)\cdot\frac{2}{3}+(ab+2a+2b+3)\cdot\frac{1}{2}$
if $b=0$:	$\binom{a}{2}\cdot\frac{3}{4}+(a+1)\cdot\frac{2}{3}+(3a+3)\cdot\frac{1}{2}$

Clearly, the case with $b \geqslant 1$ (and therefore $a \geqslant 1$) always results in bigger group eccentricity, so we assume:

$$\begin{aligned}\mathrm{GE}_1\,(G_c,C) \quad = \quad & \left(\binom{a}{2}+\binom{b}{2}\right)\cdot\frac{4}{5}+(a+b)\cdot\frac{3}{4}\\ & +(a+b+2)\cdot\frac{2}{3}+(ab+2a+2b+3)\cdot\frac{1}{2}.\end{aligned}$$

Since $a+b=n-4$ is fixed, after some calculations, the expression can be simplified to:

$$\begin{aligned}\mathrm{GE}_1(G_c,C) &= \left(\frac{a^2+b^2}{2}-\frac{n-4}{2}\right)\cdot\frac{4}{5}+(n-4)\cdot\frac{3}{4}\\ &\quad +(n-2)\cdot\frac{2}{3}+(ab+2n-5)\cdot\frac{1}{2}\\ &= \frac{2}{5}\left(a^2+\frac{5}{4}ab+b^2\right)+n\cdot\left(1+\frac{2}{3}+\frac{3}{4}-\frac{2}{5}\right)\\ &\quad -\left(\frac{5}{2}+\frac{4}{3}+3-\frac{8}{5}\right)\\ &= \frac{2}{5}\left((n-4)^2-\frac{3}{4}ab\right)+\frac{121}{60}\cdot n-\frac{157}{30}\\ &= \frac{2}{5}n^2-\frac{71}{60}n+\frac{7}{6}-\frac{3}{10}ab.\end{aligned}$$

From maximized expression it clearly follows $a = n - 5$ and $b = 1$, so we have

$$\begin{aligned}\mathrm{GE}_1(G_c, C) &= \left(\frac{(n-5)(n-6)}{2} + 0\right) \cdot \frac{4}{5} + (n-4) \cdot \frac{3}{4} \\ &\qquad + (n-2) \cdot \frac{2}{3} + (3n-10) \cdot \frac{1}{2} \\ &= \frac{2}{5}n^2 - \frac{89}{60}n + \frac{8}{3}.\end{aligned}$$

By comparing the values from all three cases above

$$\max\left(\frac{1}{3}n^2 - n + \frac{4}{3}, \frac{3}{8}n^2 - \frac{35}{24}n + 3, \frac{2}{5}n^2 - \frac{89}{60}n + \frac{8}{3}\right)$$

we can conclude that Case 3, corresponding to the rightmost graph in Fig. 7.2 is always the biggest for $n \geqslant 6$, which concludes our proof. □

We are now ready to state the following theorem.

Theorem 7.8. *For the family $\mathcal{G}_n$ and $|C| = 2$, the following pairs (G, C) are extremal:*

- *if $n = 3$, then $G = P_3$ with any choice of C;*
- *if $n = 4$, then $G = S_3$ where the center of the star is in C;*
- *if $n = 5$, then $G = D(5, 3)$ whenever C dominates G;*
- *if $n \geqslant 6$, then $G = D(n, 5)$ with $C = \{p_0, p_3\}$.*

Proof. In order to find an extremal pair in $\mathcal{G}_n$, it is by Lemma 7.4 enough to consider members of $\mathcal{T}_n$ only. We start with the small values of n. Since $n > c$, we start with $n = 3$, where the only tree to consider is P_3. It holds $\mathrm{GE}_1(G, C) = 0$ with any selection of C. For $n = 4$, we have two possible trees (see two leftmost figures in Fig. 7.3). In the left case, $\mathrm{GE}_1(S_3, C) = \frac{3}{2}$ and in the right case, $\mathrm{GE}_1(P_4, C) = 1$, making S_3 the only extremal graph on four vertices.

For $n = 5$, we have three possible trees (see three rightmost figures in Fig. 7.3). In the left case, $\mathrm{GE}_1(S_4, C) = 3$, in the middle graph $D(5, 3)$, we have $\mathrm{GE}_1(D(5, 3), C) = \frac{25}{6}$ and in the right case, it holds $\mathrm{GE}_1(P_5, C) = \frac{23}{6}$, making $D(5, 3)$ the only extremal graph on five vertices.

Now assume that $n \geqslant 6$. We will denote $C = \{x, y\}$ as maximizing pair of vertices in a graph and let $d := d(x, y)$. By Lemma 7.7, we can assume that $d(x, y) = 3$. As stated in the same lemma, Case $d(x, y) = 3$, maximum with eccentricity centralization equal to $\frac{2}{5}n^2 - \frac{89}{60}n + \frac{8}{3}$ is obtained by the graph $D(n, 5)$.

Table 7.1 shows extremal values of group eccentricity for $c = 2$, and the corresponding graphs from $\mathcal{T}_n$ (depending on the number of vertices n).

□

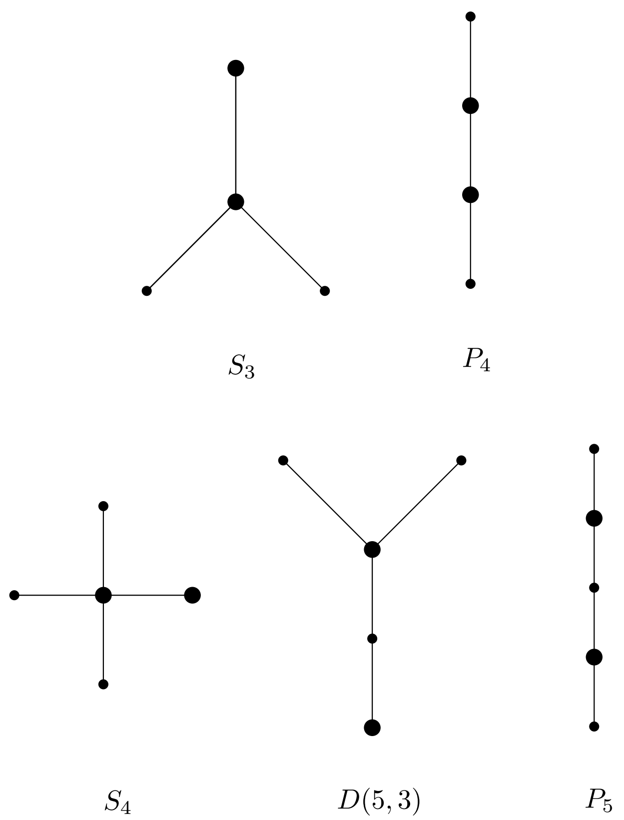


Figure 7.3: Two possible non-isomorphic trees on four vertices (top row) and three possible non-isomorphic trees on five vertices (bottom row). Members of C are emphasized.

GRAPH SIZE	AN EXTREMAL GRAPH G	MAXIMUM GROUP ECCENTRICITY
3	P_3	0
4	S_3	3/2
5	$D(5,3)$	25/6
$n \geqslant 6$	$D(n,5)$	$\frac{2}{5}n^2 - \frac{89}{60}n + \frac{8}{3}$

Table 7.1: Extremal values of group eccentricity for $c = 2$, and the corresponding graphs from $\mathcal{T}_n$.

Note that Theorem 7.8 does not characterize all the extremal graphs from $\mathcal{G}_n$, but only lists these from $\mathcal{T}_n$.

7.5 GROUP BETWEENNESS CENTRALIZATION

Everett and Borgatti [46] introduced group betweenness in the following way. Let G be a graph and $C \subseteq V(G)$. Let $\sigma_{u,v}$ be the number of geodesics connecting u to v and $\sigma_{u,v}(C)$ be the number of geodesics connecting u to v passing through some vertex of C. Then, the group betweenness centrality of C is given by

$$\mathrm{GB}(G,C) = \sum_{\{u,v\} \subseteq V(G) \setminus C} \frac{\sigma_{u,v}(C)}{\sigma_{u,v}}.$$

For easier notation, we will sometimes write $\mathrm{GB}(C)$ instead of $\mathrm{GB}(G,C)$ (where G is fixed) and $\mathrm{GB}(G)$ instead of $\mathrm{GB}(G,C)$ (where we assume that C is the maximizing set among $\binom{V(G)}{c}$, for fixed size of group c). Note that it always holds

$$\max_{v \in C} \mathrm{GB}(\{v\}) \leqslant \mathrm{GB}(C) \leqslant \sum_{v \in C} \mathrm{GB}(\{v\}). \tag{7.6}$$

In this section, we extend group betweenness with the notion of group betweenness centralization, defined in [52] and observe some of extremal-pairs, given fixed network and group sizes.

Consider the star S_{n-1} on n vertices, and let the vertex v be its center. The betweenness centrality of any leaf is zero, therefore by (7.6) group centrality of any $C \subset S_{n-1}$ is given by:

$$\mathrm{GB}(S_{n-1}, C) = \begin{cases} \binom{n-c}{2} & v \in C; \\ 0 & v \notin C. \end{cases} \tag{7.7}$$

We will now calculate the group betweenness centralization for the group C in the graph S_{n-1}. Observe, that

$$\begin{aligned} \mathrm{GB}_1(S_{n-1}) &= \binom{n}{c} \cdot \binom{n-c}{2} - \binom{n-1}{c-1} \cdot \binom{n-c}{2} \\ &= \binom{n-c}{2} \cdot \binom{n-1}{c}. \end{aligned} \tag{7.8}$$

Clearly, the expression above is a lower bound on maximum group betweenness centralization in $\mathcal{G}_n$ for given c. As we will see below, this bound is not the best.

Definition 7.9. Let $S_{n,i}$ be a graph on n vertices defined in the following way. Consider a graph $K_{n-i,i}$, and denote a bipartition of size i with I. Then

$$V(S_{n,i}) = V(K_{i,n-i}) \quad \text{and} \quad E(S_{n,i}) = E(K_{i,n-i}) \cup \{uv \mid u,v \in I\}.$$

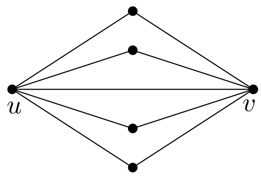


Figure 7.4: Graph $S_{6,2}$.

An example of $S_{n,i}$ is shown in Fig. 7.4. Later we will calculate that for such a graph there exists a group with higher group betweenness centralization than any group in a star graph of the same size.

Proposition 7.10. *The group betweenness centralization of $S_{n,c}$ is given by*

$$\mathrm{GB}_1(S_{n,c}) = \binom{n}{c}\binom{n-c}{2} - \sum_{j=0}^{c}\binom{c-1}{j}\binom{n-c}{j}\binom{n-c-j}{2}.$$

Proof. For simplicity, let $G := S_{n,c}$, let $\overline{C}$ be the largest independent set in $V(G)$, and let $C := V(S_{n,c}) \setminus \overline{C}$. By definition, it holds

$$\begin{aligned}\mathrm{GB}_1(G,C) &= \binom{n}{c}\mathrm{GB}(G,C) - \sum_{C' \in \binom{V(G)}{c}} \mathrm{GB}(G,C') \\ &= \binom{n}{c}\binom{n-c}{2} - \sum_{C' \in \binom{V(G)}{c}} \mathrm{GB}(G,C').\end{aligned}$$

Any $C' \in \binom{V(G)}{c}$ may have some members in C, and others in $\overline{C}$. Let $j := |C' \cap \overline{C}|$ and so $c - j = |C' \cap C|$. We will calculate $\mathrm{GB}(G,C')$. Notice that (a,b)-paths, with $ab \in E(G)$ are trivial, and do not visit any vertex along the way, therefore it is enough to consider only all (a,b)-paths for $a, b \in \overline{C}$. For every chosen pair a, b, we have $\sigma_{a,b} = c$ and $\sigma_{a,b}(C') = c - j$. Since there is precisely $\binom{n-c-j}{2}$ such pairs, we may conclude

$$\mathrm{GB}(G,C') = \binom{n-c-j}{2} \cdot \frac{c-j}{c}.$$

Since there are $\binom{n-c}{j}$ choices for a j-subset in $\overline{C}$ and $\binom{c}{c-j}$ choices for a $(c-j)$-subset in C, we have

$$\begin{aligned}\mathrm{GB}_1(G,C) &= \binom{n}{c}\binom{n-c}{2} - \sum_{j=0}^{c}\binom{c}{c-j}\binom{n-c}{j}\binom{n-c-j}{2} \cdot \frac{c-j}{c} \\ &= \binom{n}{c}\binom{n-c}{2} - \sum_{j=0}^{c}\binom{c-1}{j}\binom{n-c}{j}\binom{n-c-j}{2},\end{aligned}$$

concluding our proof. □

Before we move on to the main theorem, we introduce the following triangle-like property.

Proposition 7.11. *Let G be any fixed graph on n vertices. For any three vertices $x, y, z \in V(G)$, the following bound holds*

$$\mathrm{GB}(\{x,z\}) + \mathrm{GB}(\{z,y\}) \geqslant \mathrm{GB}(\{x,y\}) - n + 3. \tag{7.9}$$

Proof. We show that $X \geqslant 0$ where $X := \mathrm{GB}(\{x,z\}) + \mathrm{GB}(\{z,y\}) + n - 3 - \mathrm{GB}(\{x,y\})$. Notice that

$$\begin{aligned}
X = &\sum_{a,b\in V(G)\setminus\{x,y,z\}} \frac{\sigma_{a,b}(x,z)+\sigma_{a,b}(z,y)-\sigma_{a,b}(x,y)}{\sigma_{a,b}} + \sum_{a\in V(G)\setminus\{x,z\}} \frac{\sigma_{a,y}(x,z)}{\sigma_{a,y}} \\
&+ \sum_{a\in V(G)\setminus\{z,y\}} \frac{\sigma_{a,x}(z,y)}{\sigma_{a,x}} + n - 3 - \sum_{a\in V(G)\setminus\{x,y\}} \frac{\sigma_{a,z}(x,y)}{\sigma_{a,z}} \\
\geqslant &\sum_{a,b\in V(G)\setminus\{x,y,z\}} \frac{\sigma_{a,b}(x,z)+\sigma_{a,b}(z,y)-\sigma_{a,b}(x,y)}{\sigma_{a,b}} + n - 3 - \sum_{a\in V(G)\setminus\{x,y\}} \frac{\sigma_{a,z}(x,y)}{\sigma_{a,z}}.
\end{aligned}$$

Clearly, $\sigma_{a,b}(x,z) + \sigma_{a,b}(z,y) - \sigma_{a,b}(x,y)$ is always positive for any distinct vertices a, b, x, y, z, therefore first sum is always positive. The rightmost sum is at most $n-3$, since

$$\sum_{a\in V(G)\setminus\{x,y,z\}} \frac{\sigma_{a,z}(x,y)}{\sigma_{a,z}} \leqslant \sum_{a\in V(G)\setminus\{x,y,z\}} 1 \leqslant n-3.$$

Putting everything together, we can conclude

$$\begin{aligned}
X = &\sum_{a,b\in V(G)\setminus\{x,y,z\}} \frac{\sigma_{a,b}(x,z)+\sigma_{a,b}(z,y)-\sigma_{a,b}(x,y)}{\sigma_{a,b}} \\
&\qquad + \sum_{a\in V(G)\setminus\{xzy\}} \left(\frac{\sigma_{a,x}(z,y)}{\sigma_{a,x}} + \frac{\sigma_{a,y}(x,z)}{\sigma_{a,y}}\right) \\
&\qquad\qquad + n - 3 - \sum_{a\in V(G)\setminus\{zxy\}} \frac{\sigma_{a,z}(x,y)}{\sigma_{a,z}} \\
\geqslant &\, 0 + \sum_{a\in V(G)\setminus\{x,y,z\}} \left(\frac{\sigma_{a,x}(z,y)}{\sigma_{a,x}} + \frac{\sigma_{a,y}(x,z)}{\sigma_{a,y}}\right) + 0 \\
\geqslant &\, 0.
\end{aligned}$$

□

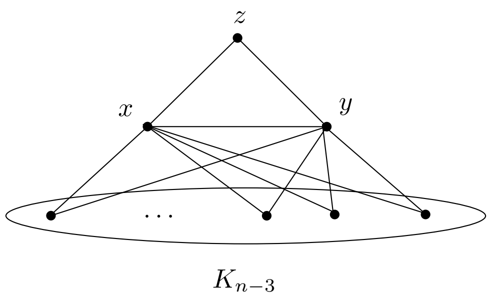


Figure 7.5: By easy observation, it holds $\mathrm{GB}(\{x,y\}) = n-3$ and $\mathrm{GB}(\{x,z\}) = \mathrm{GB}(\{z,y\}) = 0$. Therefore $\mathrm{GB}(\{x,y\}) = \mathrm{GB}(\{x,z\}) + \mathrm{GB}(\{z,y\}) + n-3$. The bound from Lemma 7.11 is tight.

To show that the bound from Proposition 7.11 is tight, we give an example of a graph on n vertices obtained as follows. Start with the complete graph K_{n-1} and denote two of its vertices by x, y. Add a new vertex z to a graph, connecting it to x and y (see Fig. 7.5). In such an arrangement, it holds $\mathrm{GB}(\{x,y\}) = n-3$ and $\mathrm{GB}(\{x,z\}) = \mathrm{GB}(\{z,y\}) = 0$. Therefore $\mathrm{GB}(\{x,y\}) = \mathrm{GB}(\{x,z\}) + \mathrm{GB}(\{z,y\}) + n-3$ and the bound from Proposition 7.11 is tight.

Now, we are ready to give the main theorem for the group betweenness centralization.

Theorem 7.12. *Let (G, C) be an extremal-pair for the graph family $\mathcal{G}_n$ with $c = 2$. Then,*

$$\mathrm{GB}_1(G,C) = \binom{n-2}{2}\left(\binom{n-1}{2} + 2\right).$$

Furthermore, the extremal value is reached at pair $(S_{n,2}, I)$.

Proof. The expression (7.8) gives us a trivial lower bound $\mathrm{GB}_1(C,G) \geqslant \binom{n-1}{2}\binom{n-2}{2}$, but we need a better lower bound. We start by showing $\mathrm{GB}_1(G,C) \geqslant \binom{n-1}{2}\binom{n-2}{2} + n^2 - 5n + 6$.

Assuming that our group is of size 2, from Proposition 7.10 it clearly follows

$$\mathrm{GB}_1(S_{n,2}, I) = \binom{n-2}{2}\left(\binom{n-1}{2} + 2\right),$$

therefore for any extremal-pair (G, C), it holds

$$\mathrm{GB}_1(G,C) \geqslant \binom{n-2}{2}\left(\binom{n-1}{2} + 2\right).$$

It remains to show, that no other pair can achieve a better score. Let (G, C) be an extremal-pair for the graph family $\mathcal{G}_n$, with $c = 2$ and $C = \{u, v\}$. Now, we maximize the possible value of the expression

$$\mathrm{GB}_1(G,C) = \binom{n}{2} \cdot \mathrm{GB}(G, \{u,v\}) - \sum_{C' \in \binom{V(G)}{2}} \mathrm{GB}(G, C').$$

To maximize $\binom{n}{2} \cdot \mathrm{GB}(G, \{u, v\})$, note that in the extremal case, all shortest paths between all pairs of vertices visit a member of C, see (7.7). Therefore $\mathrm{GB}(C, G) \leqslant \binom{n-c}{2}$ for any chosen set C. To minimize $\sum_{C' \in \binom{V(G)}{2}} \mathrm{GB}(G, C')$, we use Proposition 7.11.

Joining both conclusions, we get

$$\begin{aligned}
\mathrm{GB}_1(G, C) = \binom{n}{2} \cdot \mathrm{GB}(G, C) &- \sum_{a \in V(G) \setminus \{u, v\}} \left(\mathrm{GB}\left(G, \{u, a\}\right) + \mathrm{GB}\left(G, \{v, a\}\right)\right) \\
&- \mathrm{GB}\left(G, C\right) - \sum_{C' \in \binom{V(G) \setminus \{u,v\}}{2}} \mathrm{GB}\left(G, C'\right) \\
&\leqslant \binom{n}{2} \cdot \mathrm{GB}(G, C) - (n-2) \cdot (\mathrm{GB}(G, C) - n + 3) - \mathrm{GB}(G, C) \\
&\leqslant \mathrm{GB}(G, C) \cdot \left(\binom{n}{2} - 1 - n + 2\right) + (n-2)(n-3) \\
&\leqslant \binom{n-1}{2}\binom{n-2}{2} + n^2 - 5n + 6 \\
&= \binom{n-2}{2}\left(\binom{n-1}{2} + 2\right),
\end{aligned}$$

which concludes the proof. □

7.6 SOME OPEN PROBLEMS

We determined that the star is the maximizing graph for degree centralization, which is not surprising. One would think that also for other centralization measures, the star would be the best possible choice. To our surprise, this is not the case. The maximum achieved value of group eccentricity centralization for 2-sets is realized by Dandelion graphs. Stars are also not extremal graph for betweenness centralization for $c = 2$. We give the following conjectures about the structure of extremal graphs.

Conjecture 7.13. *$(S_{n,c}, C)$ is the extremal-pair for group betweenness centralization of networks on n vertices and groups of size $|C| = c$.*

Our next problem is regarding the eccentricity centralization. In particular, we believe that the graphs of extremal-pairs are Dandelion graphs.

Conjecture 7.14. *Let $c \leqslant \frac{n}{3}$ and let $C = \{p_{3i}\}_{i=0}^{c-1}$, where p_i are vertices of $D(n, 3c-1)$, as defined in Section 7.4. Then $(D(n, 3c-1), C)$ is an extremal-pair for group eccentricity centralization of networks on n vertices.*

Another interesting question for studying centrality indices is presented in [48], where authors are asking, in the class of graphs with fixed maximum degree and fixed number of edges, for any extremal-pair for betweenness centralization (G, v) it holds $\deg(v) = \Delta(G)$. As presented in this chapter, the mentioned conjecture holds also for eccentricity group centralization on 2-sets, and for degree group centralization. It is therefore natural to ask the following:

Conjecture 7.15. *Let (G, C) be an extremal-pair for group eccentricity centralization of networks on n vertices, with arbitrary size of set C. Then, there exists a member $v \in C$, such that* $\deg(v) = \Delta(G)$.

8

ALGORITHMIC APPROACH TO DEGREE CENTRALIZATION IN LARGE NETWORKS

In the thesis, much has been written on centrality indices, Freeman centralization, and group centrality concepts, see Section 3.3 on page 21, Section 3.3.2 on page 24 and Chapter 7 on page 89, respectively. The most basic centrality measure of *degree centrality* is simply defined as the degree of a given node. In this chapter, we study group centralization notion for degree centrality, i.e. group degree centralization.

The chapter is structured as follows. In Section 8.1 we provide notations and definitions from [84] that we use. In Section 8.2 we mathematically analyze group degree centralization problem and determine its time-complexity. Furthermore, using a classic graph theoretical approach of double counting, we show that the sum $\sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G)$ can be computed efficiently. In Section 8.3, with help of a related problem k-MaxVD (see Miyano and Ono [107]) we develop an efficient greedy algorithm for finding an approximate group with maximal degree centralization. The algorithm approximately determines group degree centralization for all group sizes k (with $1 \leqslant k \leqslant n$) and altogether runs in $O(m + \gamma(G) \cdot \Delta(G)) \leqslant O(n^2)$ time. We describe the procedure in detail and provide complexity analysis of the algorithm. In Section 8.4 we describe the experiments made on six real-world networks. We present computers and datasets used and discuss the results. In the results we observe the *degree-unimodality* property, which may be a new property for studying real-world networks. In the concluding section we provide some challenges for possible future work.

8.1 PRELIMINARIES

Let G be a graph on n vertices and m edges, and let $S \subseteq V(G)$. According to [46], group degree centrality is defined as

$$\mathrm{GD}_G(S) = \left| \bigcup_{v \in S} N(v) \setminus S \right|.$$

For a given k, let S_k^* be one of the sets from $\binom{V(G)}{k}$ that achieves the maximum value of group degree centrality, i.e. $\mathrm{GD}_G(S_k^*) = \max_{S\in\binom{V(G)}{k}} \mathrm{GD}_G(S)$.

According to [52, 84], $\mathrm{GD}_1(S, G)$ stands for group degree centralization. Define $k := |S|$, and observe

$$\mathrm{GD}_1(S,G) \quad = \quad \frac{\sum_{S'\in\binom{V(G)}{k}} (\mathrm{GD}_G(S) - \mathrm{GD}_G(S'))}{\max_{H\in\mathcal{G}_n} \sum_{S''\in\binom{V(H)}{k}} \mathrm{GD}_H(S^*_{|S|}) - \mathrm{GD}(S'')}. \tag{8.1}$$

According to Freeman [52], the denominator is needed to efficiently normalize centralization to interval $[0, 1]$, for better relative comparison. Clearly $\mathrm{GD}_1(S, G)$ is maximized whenever $\mathrm{GD}(S, G)$ is maximized, and from [84] we know that the maximum value of the denominator corresponds to star graph S_n, where an optimal set S_k^* is any set containing the center of the star.

Whenever the graph G is known from the context, we omit it from the notions of centrality or centralization. Denote *maximizing group size* $\mathrm{dc}(G)$ to be the positive integer such that $S^*_{\mathrm{dc}(G)}$ achieves the maximum value of group degree centralization, i.e. $\mathrm{GD}_1(S^*_{\mathrm{dc}(G)}, G) = \max_{k\in[n]} \mathrm{GD}_1(S_k^*, G)$, and also denote $S^* := S^*_{\mathrm{dc}(G)}$. Let $\gamma(G)$ be a cardinality of a minimum set that dominates graph G (also known as *domination number*). The notation $\Delta(G)$ stands for the highest degree of any vertex in a graph G, i.e. $\Delta(G) = \max_{v\in V(G)} \deg_G(v)$. A function f is said to be *unimodal* if locally there is only a single highest value in f.

8.2 EVALUATING DEGREE CENTRALIZATION

The goal of this section is to optimize the procedure of calculating group degree centralization for a given graph and a constant k. We start by calculating the denominator for group degree centralization.

Proposition 8.1. *Let G be a star on n vertices with center c, and let $S^* \in \binom{V(G)}{k}$ such that $c \in S^*$. Then*

$$\sum_{S'\in\binom{V(G)}{k}} \left[\mathrm{GD}(S^*) - \mathrm{GD}(S')\right] = (k+1)\binom{n-1}{k+1}.$$

Proof. Let us partition the sets of $\binom{V(G)}{k}$ into two parts P_1 and P_2, depending on weather or not (respectively) they include the vertex c as a member. It is easy to see that the group degree centrality of members of these parts equals to $n - k$ and 1, respectively.

Consider the number of k-sets of $V(G)$ that contain the vertex c, i.e. $|P_1|$. Since all such sets contain vertex c, we have yet to choose $k - 1$ set members among the remaining $n-1$ vertices, therefore we have $|P_1| = \binom{n-1}{k-1}$. With the similar argument, we get $|P_2| = \binom{n-1}{k}$. Joining those facts, we get

$$\begin{aligned} & \sum_{S' \in \binom{V(G)}{k}} \left[\mathrm{GD}(S^*, G) - \mathrm{GD}(S', G)\right] \\ = & \binom{n}{k}(n-k) - \binom{n-1}{k-1}(n-k) - \binom{n-1}{k} \\ = & (n-k) \cdot \binom{n-1}{k} - \binom{n-1}{k} \\ = & (n-k-1)\binom{n-1}{k} \\ = & (k+1) \cdot \binom{n-1}{k+1}. \end{aligned}$$

□

In the following proposition we use a classic graph theoretical approach of double counting to show, that the sum $\sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G)$ from (8.1) can be computed efficiently.

Proposition 8.2. *Let G be a graph on n vertices, and let $k \leqslant n$ be a positive integer. It holds that*

$$\sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G) = n \cdot \binom{n-1}{k} - \sum_{v \in V(G)} \binom{n - \deg(v) - 1}{k}.$$

In particular, the sum $\sum_{S' \in \binom{V(G)}{k}} \mathrm{GC}(S', G)$ can be computed in $O(n)$ steps.

Proof. For each vertex $v \in V(G)$, define its contribution $g_k(v)$ to be the number of k-sets that dominates v, i.e.

$$\begin{aligned} g_k(v) &= \left| \left\{ X \in \binom{V(G) - v}{k}; X \cap N(v) \neq \emptyset \right\} \right| \\ &= \binom{n-1}{k} - \binom{n - \deg(v) - 1}{k}. \end{aligned}$$

It follows that

$$\begin{aligned} \sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G) &= \sum_{v \in V(G)} g_k(v) \\ &= n \cdot \binom{n-1}{k} - \sum_{v \in V(G)} \binom{n - \deg(v) - 1}{k}, \end{aligned} \tag{8.2}$$

which can be computed in $O(n)$ steps, traversing all vertices once. □

We join results from Propositions 8.1 and 8.2 to further develop (8.1). We can therefore claim the following.

Theorem 8.3. *For a given graph G and a group of its vertices S of size k, group degree centralization can be evaluated as*

$$\mathrm{GD}_1(S,G) = \frac{\binom{n}{k}\cdot \mathrm{GD}(S,G) + \sum_{v\in V(G)}\binom{n-\deg(v)-1}{k} - n\cdot\binom{n-1}{k}}{(k+1)\cdot\binom{n-1}{k+1}}, \tag{8.3}$$

which can be computed in $O(n)$ steps.

It is easy to see that finding the best-possible group S^* that maximizes group degree centrality (and hence group degree centralization) is $\mathcal{NP}$-hard.

Proposition 8.4. *Suppose that you are given an input graph G and integer k. The problem that determines a set S_k^* is $\mathcal{NP}$-hard.*

Proof. We prove the claim by reducing the problem of finding S_k^* to the existence of k-dominating set. Let us assume that there exists of polynomial algorithm for finding a k-set $S_k^* \subseteq V(G)$ such that

$$\mathrm{GD}_G\left(S_k^*\right) = \max_{S\in\binom{V(G)}{k}} \mathrm{GD}_G\left(S\right).$$

Now observe that the existence of k-dominating set is equivalent to the property

$$\mathrm{GD}_G\left(S_k^*\right) = n-k.$$

As group degree centrality of a given fixed set S_k^* can be computed in at most linear time, it is clear that the set S_k^* provides us with an answer of the existence of k-dominating set, which is a well-known $\mathcal{NP}$-hard problem. □

In Section 8.4 we present an efficient algorithm that achieves the best possible linear-time approximation for calculating group degree centrality scores for all group sizes.

8.3 ALGORITHMIC APPROACH

In this section we will present methods used to calculate an estimate of group degree centralization for big real-world networks. We introduce a greedy algorithm for finding an approximate group with maximal degree centralization, describe it in detail and provide its complexity analysis. In the end, we briefly discuss the approximation ratio.

8.3.1 *The algorithm*

Finding the group with biggest degree centralization can be useful for many real-world networks which can in some cases be very large (largest network from the experiments contains 16 million edges). To calculate group degree centralization, we reduce it to k-*VertexMaximumDomination* (k-MaxVD, for short). The problem k-MaxVD is a special case of *MaximumCoverageProblem*, that for a positive integer k finds a subset of size k that maximizes the cardinality of dominated vertices. By Proposition 8.4, group degree centralization cannot be calculated in polynomial time, unless $\mathcal{P} = \mathcal{NP}$. To calculate it efficiently, we use greedy algorithm for k-MaxVD from [107] which is a polynomial time $(1 - 1/e)$-approximation algorithm. Furthermore, in [107] it is proven that no other polynomial time constant factor approximation algorithm for k-MaxVD can have an approximation ratio better than $(1 - 1/e)$. An implementation of procedure that calculates an approximation for the group with biggest degree centralization, for all meaningful group sizes, is given by Algorithm 8.1. Note that by data structure of a *dictionary* D we mean that D is a set of keys with additionally defined values $D[i]$ for each $i \in D$. In next paragraph we describe the parts of the algorithm without going to detail; some details are included in the paragraphs that follow later.

In the first phase of Algorithm 8.1 (throughout lines 1–12), we pre-process our graph G', convert it to directed graph G, and initialize the starting values of all dictionaries and other variables. In the *main loop* (lines 14–28) we use k-MaxVD greedy approach to efficiently find groups S, starting with $k = 1$ and increasing the group size until the graph is dominated. We first choose the vertex v to add to our group S, and accordingly update variables S and k (lines 14–15). Then we calculate the group degree centrality for increased set S (in line 16). The actual changes in graph G are made throughout lines 18–25, where we remove v from G, remove the in-edges attached to the out-neighbors of v (since they just became dominated by v), and update the dictionary *histogram* accordingly. Finally, in lines 26 and 27, we update some centralization variables and calculate the Freeman centralization of the centrality from line 16. We now present some details of how we maintain some values and graph properties.

Every time that k increases, we add some greedily chosen vertex (the one that maximizes the contribution) to the set S and remove it from G while maintaining some dictionaries that we use (*contribution*, *dominated* and *histogram*, in particular). While the directed graph G that we work with is changing with each iteration, notice that the original instance of the original graph G' stays the same throughout an algorithm. Note that we initially have $\deg_G^+(v) = \deg_G^-(v) = \deg_{G'}(v)$ for all vertices from a network. To calculate Freeman centralization, we efficiently calculate group degree centralization of all possible

Algorithm 8.1 Finding group with maximal degree centralization

Input: a graph G'.

Output: a list *centralization* of group degree centralization scores, where *centralization*$[i]$ is an approximation of $\max_{S \in \binom{V(G)}{i}} \mathrm{GD}(S, G')$.

```
n ← |V(G')|, k ← 0, S ← ∅                          ▷ Centrality variables initialization.
dominated ← ∅, histogram ← ∅
G ← {a directed instance of graph G'}
for all v in V(G') do
    add v to histogram[deg_G'(v)]
    dominated[v] ← False
centralization ← ∅                                  ▷ Centralization variables initialization.
A ← 1/(n − 1)
C ← n/(n − 1)
for all i ∈ histogram do
    sum[i] ← 1/(n − 1)
    degDistribution[i] ← |histogram[i]|
while max(histogram) ⩾ 0 do                         ▷ Main loop
    v ← any vertex from histogram with the highest contribution
    S ← S ∪ v, k ← k + 1
    centrality ← centrality+CONTRIBUTION(v)
    for all u ∈ N+_G(v) do
        DECREASECONTRIBUTION(u)
    for all u ∈ N−_G(v) do
        DECREASECONTRIBUTION(u)
        dominated[u] ← True
        for all w ∈ N+_G(u) do
            DECREASECONTRIBUTION(w)
            E(G) ← E(G) − uw
    G ← G − v
    UPDATECENTRALIZATIONVARIABLES()
    centralization[k] ← A · centrality + B − C      ▷ Computing centralization.
return centralization
```

k-sets in G' by use of Propositions 8.1 and 8.2. Before the beginning of each iteration of the *main loop*, the existence of a directed edge uv means, that

- in a initial graph G', we have $uv \in E(G')$,
- neither of v, u is a member of S, and

- in a initial graph G', vertex v is not connected with any vertex from S, i.e. $v \notin \cup_{v \in S} N(v)$.

While the first two properties are trivial to prove, the third follows from Lines 23–24 of Algorithm 8.1. For any vertex $v \in V(G) \setminus S$, we define the *contribution* of vertex v to be the value $\mathrm{GD}(S \cup \{v\}, G) - \mathrm{GD}(S, G)$ and observe that

$$\mathrm{GD}(S \cup \{v\}, G) - \mathrm{GD}(S, G) = \begin{cases} \deg_G^-(v) & \text{if } v \text{ is not dominated,} \\ \deg_G^-(v) - 1 & \text{otherwise.} \end{cases}$$

The calculation of the value contribution is done by a short function *contribution* (see Algorithm 8.3). As different nodes have various contributions we define a dictionary *histogram*, initialized in line 5 of Algorithm 8.1, where keys are all possible values of contribution (for any key i, it clearly holds $-1 \leqslant i \leqslant \Delta(G)$), while the values are unordered sets of nodes with a given contribution. While dictionary *histogram* is initially indeed a degree-histogram, with each modification of graph G and set S we carefully update it. The goal of algorithm is, for each S, to calculate the value of (8.3). We implement this by introducing variables $A, B, C, centrality$, each assigned for different part of the expression (8.3), i.e.

$$\mathrm{GD}_1(S, G) = \frac{\binom{n}{k} \cdot \mathrm{GD}(S, G) + \sum_{v \in V(G)} \binom{n-\deg(v)-1}{k} - n \cdot \binom{n-1}{k}}{(k+1) \cdot \binom{n-1}{k+1}},$$

which may, for algorithmic purposes, be written as $A \cdot centrality + B - C$, with

$$\begin{aligned} A &= \frac{n}{(n-k)(n-k-1)}, \\ centrality &= \mathrm{GD}(S, G), \\ B &= \frac{(n-k-2)!}{(n-1)!} \sum_{v \in V(G)} \frac{(n-\deg(v)-1)!}{(n-k-\deg(v)-1)!}, \\ C &= \frac{n}{n-k-1}. \end{aligned}$$

Clearly, whenever the group S increases, also all values of $A, B, C, centrality$ change. To handle variable B we also introduce a dictionaries *degDistribution* and *sum*, where keys of both are all degrees of vertices in G, and the values are defined as

$$\begin{aligned} sum[i] &= \frac{(n-i-1)!}{(n-k-i-1)!} \cdot \frac{(n-k-2)!}{(n-1)!}, \\ degDistribution[i] &= \text{number of vertices in } G \text{ of degree } i. \end{aligned}$$

Algorithm 8.2 Updating variables A, B, C and *sum*.

1: **function** UPDATECENTRALIZATIONVARIABLES
2: $\quad A \leftarrow \frac{n}{(n-k)(n-k-1)}$
3: $\quad C \leftarrow \frac{n}{n-k-1}$
4: $\quad$ **for all** i in *sum* **do**
5: $\quad\quad sum[i] \leftarrow sum[i] \cdot \frac{N-i-k-1}{N-k-2}$
6: $\quad B \leftarrow \sum_i sum[i] \cdot degDistribution[i]$

Note that *sum* need to be refreshed with every change of k. We first initialize both dictionaries before entering the main loop, and then we maintain their values by use of a function *updateCentralizationVariables* (we treat these variables as global variables, therefore no parameters are needed). To avoid using big numbers, we update the value of $sum[i]$ by just multiplying it by $\frac{(n-k-i-1)}{(n-k-2)}$, whenever k increases by one. Using this, B can be calculated by a simple addition

$$B = \sum_i sum[i] \cdot degDistribution[i],$$

see line 6 of Algorithm 8.2.

Note that while the graph is changing and the group S is increasing, also contributions of the remaining vertices change. While initial contribution for some vertex v is equal its degree $\deg_G(v)$, during the main loop the contribution of v may incrementally decrease by one several times. We handle these changes by defining a function $decreaseContribution(v)$, see Algorithm 8.3.

Algorithm 8.3 Functions *contribution* and *decreaseContribution*. The former outputs the contribution of v while the latter refreshes the dictionary *histogram* whenever the contribution of node v decreases by one.

function CONTRIBUTION(v)
$\quad$ **if** $dominated[v]$ **then**
$\quad\quad$ **return** $\deg^-_G(v) - 1$
$\quad$ **else**
$\quad\quad$ **return** $\deg^-_G(v)$

function DECREASECONTRIBUTION(v)
$\quad c \leftarrow$ CONTRIBUTION(v)
$\quad histogram[c] \leftarrow histogram[c] \setminus \{v\}$
$\quad histogram[c-1] \leftarrow histogram[c-1] \cup \{v\}$

8.3.2 *Complexity analysis*

In this section, we analyze the running time complexity of Algorithm 8.1 in terms of n, m, $\Delta(G)$ and $\gamma'(G) := \frac{e}{e-1}\gamma(G)$. First notice that both functions from Algorithms 8.2 and 8.3 take at most $O(\Delta(G)) \leqslant O(n)$ and $O(1)$, respectively. Using the greedy approach from [107], in each step of the algorithm a node with biggest contribution to the group degree centrality is added to the group. We implement this by maintaining a sorted histogram of node contributions. Initially, the contributions to group centrality are equal to node degrees. After each addition of a new node to the group, the histogram is accordingly updated. The *MAIN LOOP* of the algorithm terminates when there is no node with contribution greater than -1, which occurs when all the nodes are dominated by the group. Note that the the number of iterations of the *MAIN LOOP* is bounded to $\gamma'(G)$.

It is easy to see, that throughout lines 1–10, our algorithm needs up to $O(n)$ steps to initialize all needed variables for the *MAIN LOOP* to start. By use of the dictionary *histogram*, it takes constant time to greedily pick the vertex v with biggest contribution, therefore line 12 altogether takes at most $O(n)$ time. Further notice that also lines 13–15 and 25 take constant time, altogether using up to additional $O(n)$ steps. Consider now the lines 16–23 and notice, that a function *decreaseContribution* is called for each removed directed edge precisely once (here we also consider edges removed in line 23. Furthermore, in line 22, deleting a vertex v from a graph means removing a vertex ($O(1)$) and removing its adjacent edges ($O(\deg(v))$). Since there is m edges in G, the lines 16–23 are altogether bounded to $O(m)$. Finally, line 24 takes constant time for updating variables A and C, while both *sum* and B takes up to $O(\Delta(G))$ time to compute, altogether summing up to $O(\gamma'(G) \cdot \Delta(G))$. Summing everything, the final complexity of finding the approximation of group degree centralization for groups of sizes $1 \leqslant k \leqslant n$ in Algorithm 8.1 equals $O(m + \gamma(G) \cdot \Delta(G)) \leqslant O(n^2)$. For most scale-free networks this can probably be reduced further to $O(n \log n)$. From [107] we know that from approximation ratio point of view, Algorithm 8.1 is the best possible polynomial approximation for group degree centralization.

8.3.3 *Performance guarantee*

In this section we determine the absolute performance guarantee of out algorithm. Let S^* be a set that maximizes group degree centrality in G, and let S be our sub-optimal set, found by greedy algorithm from [107] with relative performance guarantee $1 - \frac{1}{e} \approx 0.632$. Based on such low performance guarantee, we do not expect our algorithm to perform well theoretically. Although the approximation ratio from [107] is quite bad, authors prove that it is the best that one can attain in polynomial time.

Let us now focus on Algorithm 8.1. From expression (8.3), it follows

$$\begin{aligned}
\mathrm{GD}_1(S^*,G) &\leqslant \frac{\binom{n}{k}\cdot\frac{e}{e-1}\mathrm{GD}(S,G)+\sum_{v\in V(G)}\binom{n-\deg(v)-1}{k}-n\cdot\binom{n-1}{k}}{(k+1)\cdot\binom{n-1}{k+1}} \\
&= \mathrm{GD}_1(S,G)+\frac{n\cdot\mathrm{GD}(S,G)}{(e-1)\,(n-k)\,(n-k-1)} \\
&\leqslant \mathrm{GD}_1(S,G)+\frac{n}{(e-1)\,(n-k-1)}.
\end{aligned}$$

Thus, the error of our approximation algorithm is bounded to $\frac{n}{(e-1)(n-k-1)}$. In the table 8.1, the error bound is calculated for the maximal achieved group degree centralization value in the datasets we used.

G	$\lvert V(G)\rvert$	$\mathrm{GD}_1(S^*_{\mathrm{dc}(G)},G)$	ERROR BOUND
facebook	4,039	0.905393	0.5836
cobiss	25,301	0.654997	0.4590
twitter	81,306	0.773615	0.5754
amazon	403,394	0.542647	0.4926
youtube	1,134,890	0.777639	0.7172
patents	3,923,922	0.470009	0.5925

Table 8.1: The theoretical error-bound for the experiments we did.

8.4 EXPERIMENTS

In this section we describe the experiments made with six real-world networks. We present the machines and datasets used and discuss the results. In the results we observe the *degree-unimodality* property, which may be a new property for studying real-world networks.

8.4.1 *Datasets*

Here we describe the datasets used for testing the degree centralization algorithm. We used six real world networks ranging from several hundred thousand up more than 16 million edges.

Facebook is the smallest dataset in experiments containing 4039 nodes and 88234 edges [92]. It is anonymized data collected by survey participants using a Facebook application with combined ten ego networks. The dataset was generated for study of social circles in ego networks [93]. Cobiss dataset is a graph of scientific co-authoring of the complete national research database in Slovenia from 1970 to 2013. Two authors are connected if they publish at least one paper together. The graph contains 25,301 nodes and 316,587 edges. The dataset was generated using database maintained by ARRS (Slovenian Research Agency) and IZUM (Institute of Information Science, Maribor, Slovenia). Twitter dataset contains graph of followers, with 81,306 nodes and 1,768,149 edges [92]. The dataset was collected from public sources for study of social circles discovering study [93]. Amazon dataset is a graph of frequently co-purchased products based on Amazon website in June, 2003 [92]. The graph has 403,394 nodes and 3,387,388 edges and was generated for the study of viral marketing dynamics [97]. Youtube dataset contains a graph of user friendship graph [92]. The graph contains 1,134,890 nodes and 2,987,624 edges. The dataset was prepared by Mislove et al [106]. Patents dataset is a citation graph of patents granted between 1975 and 1999 [92]. The graph contains 3,923,922 nodes (patents) and 16,522,438 edges (citations). The dataset was generated for the purpose of graph evolution study [96], using the U.S. patent dataset maintained by the National Bureau of Economic Research [65].

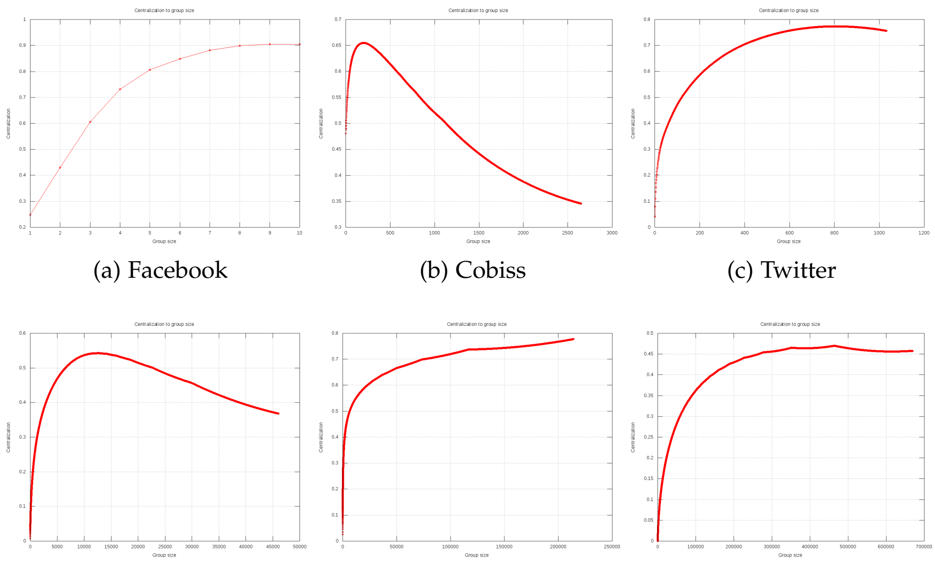
8.4.2 *Environment*

We implement the algorithm in C++[1] and Python[2] with help of the libraries "SNAP" [94] and "SNAP.Py" [95], respectively. We run the C++ implementation of the algorithm for finding group with maximal degree centralization on two different machines. First machine was a Windows Server 2012, with 4 AMD Opteron 6386 Processors (2.8 GHz), 512 GB RAM and Visual Studio 2013 environment. Second machine was a personal computer with 4 core 2.67 GHz Intel i7 CPU, 6 GB RAM, running on Ubuntu 14.04 LTS with GNU C++ compiler.

8.4.3 *Experimental results*

Figure 8.1 shows the value of centralization with different size of the group, while Table 8.2 gives precise values about results and optimal dataset sizes. In Cobiss (Figure 8.1b) network the maximal centralization is obtained with relatively small group size, and after that point adding members to the group causes drastic decreasing of centralization. Amazon (Figure 8.1d) has a similar shape, but increasing and decreasing of centralization is less intensive.

1 https://github.com/mkarlovc/gcentralization
2 https://github.com/mkrnc/group_degree_centralization



(a) Facebook (b) Cobiss (c) Twitter

(d) Amazon (e) Youtube (f) Patents

Figure 8.1: The graphic representation of Freeman centralization of group degree centrality of six networks with different sizes of groups.

In Youtube network (Figure 8.1e), relatively small group size has a high degree centralization. Further increasing of group size slowly increases centralization, which is maximal only when the group dominates all the nodes. In Facebook network (Figure 8.1a) the maximal centralization is also achieved when the group dominates all the nodes. Patents network (Figure 8.1f) is similar to Youtube, but the maximal centralization is achieved before the nodes are dominated by the group. In Twitter (Figure 8.1c), network centralization increases up to maximal point and then decreases with the same intensity.

Centralization is comparable between different networks. We can determine that the biggest centralization from the networks in the experiments obtains Facebook with values of 0.9, when the group is of size 10. This is very expectable result because the Facebook network was generated by combining 10 ego networks and the centers of the ego networks are the group members identified by our algorithm. Youtube and Twitter have relativity similar and high maximal centrality, which is obtained with 214,003 and 803 group members respectively. Cobiss and Amazon graphs are in the middle range of our centralization experiments, while the lowest maximal centrality has Patents network, which is also the largest and the sparsest network in our experiments.

Looking at the centralization values for all networks in the experiments, with thousands of different sizes of groups, we observed an interesting property, that we believe is worth further discussion. Within a small error bound, the shape of most of plots on Figure 8.1 is unimodal, i.e. it is monotonic increasing up to a maximizing group size, while at value $\mathrm{dc}(G)$ the slope starts to have negative slope. Results of our experiments indicate that this is a property that could hold for many real world networks.

G	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$	$\mathrm{dc}(G)$	$\mathrm{GD}_1(S^*_{\mathrm{dc}(G)}, G)$	$\mathrm{GD}(S^*_{\mathrm{dc}(G)}, G)$
facebook	4,039	88,234	10	0.905393	4029
cobiss	25,301	316,587	204	0.654997	19635
twitter	81,306	1,768,149	803	0.773615	78,811
amazon	403,394	3,387,388	12,810	0.542647	320,133
youtube	1,134,890	2,987,624	214003	0.777639	920,887
patents	3,923,922	16,522,438	464,298	0.470009	3,105,485

Table 8.2: Some statistics and centrality results from our experiments.

8.5 CONCLUDING REMARKS AND FUTURE WORK

Regarding group degree centralization, various results can be found in this chapter. We developed some mathematical optimizations and also used known results of Miyano and Ono [107] to develop a novel algorithm for estimating group degree centralization for all group sizes. In Section 8.4 we describe the developed algorithm and analyze its time complexity. From the theoretical performance-guarantee point of view, the algorithm may look useless, however it might be interesting to compare the attained error on real-world networks. We run the algorithm on six real-world networks on various number of nodes, ranging from few thousands to few millions and observe an interesting property of the shape of plots on Figure 8.1, which we call the *degree-unimodality* property. Out of these six networks only Patents is not unimodal although in some sense it is "almost" unimodal. Nonetheless, we believe this is a property of most of the real networks.

Finally, one may study if the unimodality property is satisfied for the group centralization of some other centrality indices, such as betweenness, closeness or eccentricity. Or, even more going out of centralization, for any "normalized" group centrality.

9

ON THE WIENER INVERSE INTERVAL PROBLEM

The Wiener index $W(G)$, introduced by Wiener [142], is among most known indices for chemical networks and is defined as the sum of the lengths of shortest paths between all pairs of vertices in G (for more motivation and background on the Wiener index the reader is referred to Section 3.4.5 on page 30).

As stated in Section 3.1, one of important aspects of network analysis is describing various network descriptors, such as degree distribution, diameter, clustering coefficient and various other structural indices. For these descriptors, we are interested in the feasible values that various networks can achieve. Many results in the field of network theory are actually of this type. For instance, a famous result of Milgram [103] claiming that people in the United States are separated by about six people on average is actually about measuring an average distance in a network. Another example of similar type by Onnela, Saramäki, Hyvönen, Szabó, Lazer, Kaski, Kertész, and Barabási [115] states that in real-world network G the value of exponent γ in the expression for expected degree $P[\deg(v) = k] = k^{-\gamma}$ of any node $v \in V(G)$ is between 2 and 3. In this chapter we are studying the values of the Wiener index in the family of all connected graphs on n nodes $\mathcal{G}_n$.

For a given integer k, the *inverse Wiener index* problem is a problem of finding a graph G, such that $W(G) = k$. The problem was proposed in 1995 by Gutman and Yeh [61], where they posed the following conjecture.

Conjecture 9.1. *For all but finitely many integers w, there exist trees with Wiener index w.*

The conjecture was first checked for integers up to 1206 by Lepović and Gutman [91], where authors found 49 integers without Wiener inverse. This result was further extended to integers up to 10^8, see Ban et al. [10]. The conjecture was finally solved in 2006, see Wagner [136] and Wang and Yu [137]. Furthermore, Fink et al. [50] showed that every sufficiently large integer has sub-exponentially many Wiener inverses in the family of trees. A related question that we deal with in this section is the following:

Problem 9.2. *What value of the Wiener index can a connected graph on n vertices have?*

In relation to this one can also ask how many such values exist, how are they distributed along the related interval or how many of them are contiguous. We will call this problem the *Wiener inverse interval problem*.

9.1 BASIC NOTIONS

Let $\mathcal{G}_n$ represent the family of all connected graphs on n vertices. For integers a and b, the notion $[a, b]$ stands for the set containing a, b and all consecutive integers between them. We now define the core notation for the Wiener inverse interval problem.

Throughout this chapter we use notion of various graph families defined in Section 2.2 on page 8, such as stars, Dandelion graphs, paths and complete graphs. Additionally, let $P(\alpha_1, \ldots, \alpha_k)$ be the graph constructed from a copy of P_k, with additional α_i leaves added to i-th vertex of a path.

Definition 9.3. For a fixed n, we define $W[\mathcal{G}_n]$ to be the image of W under $\mathcal{G}_n$, i.e.

$$i \in W[\mathcal{G}_n] \Leftrightarrow \text{ there is a graph } G \in \mathcal{G}_n \text{ such that } W(G) = i.$$

Also, let W_n^{int} be the largest interval of contiguous integers, such that $W_n^{\text{int}} \subseteq W[\mathcal{G}_n]$.

The extremals of the set $W[\mathcal{G}_n]$ are well-known, see [145].

Proposition 9.4. *For the family $\mathcal{G}_n$ it holds that*

$$\min(W[\mathcal{G}_n]) = \binom{n}{2} \text{ and } \max(W[\mathcal{G}_n]) = \binom{n+1}{3},$$

which are achieved at K_n and P_n, respectively.

A nice example is set $W[\mathcal{G}_4]$, which does not miss any value between $\binom{4}{2}$ and $\binom{5}{3}$.

Example 9.5. Let $n = 4$ and let K_4^- be a graph, isomorphic to complete graph K_4 without one edge and observe that

$$\mathcal{G}_4 = \left\{P_4, S_3, C_4, C(4,2), K_4^-, K_4\right\}.$$

In Table 9.1 the values of Wiener index for each graph from $\mathcal{G}_4$ is calculated. From Table 9.1 we deduce that $W[\mathcal{G}_4] = \{6, 7, 8, 9, 10\}$.

For the graphs with induced star, observe the following.

GRAPH FROM $\mathcal{G}_4$	THE WIENER INDEX OF A GRAPH
P_4	10
S_3	9
C_4	8
$C(4,2)$	8
K_4^-	7
K_4	6

Table 9.1: Values of Wiener index of the members of $\mathcal{G}_4$.

Observation 9.6. Let $G \in \mathcal{G}_n$ be a graph that contains k leaves adjacent to the same neighbor. Then

$$\left[W(G) - \binom{k}{2}, W(G)\right] \subseteq W\left[\mathcal{G}_n\right].$$

Proof. Let v be the common neighbor and label its adjacent leaves with $v_1, \ldots, v_k$. Now iteratively add $\binom{k}{2}$ additional edges between the leaves $v_1, v_2, \ldots, v_k$ to G, each time decreasing the Wiener index of G by one. This concludes the proof of the observation. □

The Wiener index for the family of Dandelion graphs is determined in the next lemma.

Lemma 9.7. *For a fixed positive integer n let a and b be two positive integers that sum up to n. The Wiener index of a Dandelion graph can be expressed as follows*

$$W\left(D\left(n,b\right)\right) \quad = \quad \binom{b+1}{3} + \left(\binom{b+1}{2} - 1\right) \cdot a + a^2.$$

Proof. We partition all the pairs of vertices of the graph $D(n,b)$ into three parts. First consider the pairs that belong to the path P_b. They sum up to the Wiener index of a path, i.e. $\binom{b+1}{3}$. Now consider all the pairs from the star S_a (note that this is a tree on $a+1$ nodes and a leaves). One can deduce that

$$W(S_a) = 1 \cdot a + 2 \cdot \binom{a}{2} = a^2.$$

Finally, consider all the remaining pairs. Those are all of type (u,v), where u is one of the leaves of the star, and v is one of the path-vertices, excluding the center of the star. One can easily conclude that for each of a admissible leaves of the star such distances sum up to $\binom{b+1}{2} - 1$. □

The values of Wiener index for some relevant graph families are listed below:

$$\begin{aligned} W(S_{n-1}) &= (n-1)^2 \\ W(K_n) &= \binom{n}{2} \\ W(P_n) &= \binom{n+1}{3} \\ W(C(n,b)) &= \binom{b+1}{3} + \left(\binom{b+1}{2} - 1\right) \cdot (n-b) + \binom{n-b+1}{2} \\ W(D(n,b)) &= \binom{b+1}{3} + \left(\binom{b+1}{2} - 1\right) \cdot (n-b) + (n-b)^2 . \end{aligned}$$

In the next section, we show that W_n^{int} is of length $\frac{1}{6}n^3 + O(n^2)$ and that it starts at $\binom{n}{2}$. Consequently the same holds for $W[\mathcal{G}_n]$. In the concluding section, we discuss some other properties of $W[\mathcal{G}_n]$ and W_n^{int} and state some open problems.

9.2 THE CARDINALITY OF W_n^{int} AND $W[\mathcal{G}_n]$

By Proposition 9.4, the Wiener index of every graph on n vertices falls inside the interval $\left[\binom{n}{2}, \binom{n+1}{3}\right]$. Since $\binom{n+1}{3} - \binom{n}{2} + 1 = \frac{1}{6} \cdot n^3 - \frac{1}{2} \cdot n^2 + \frac{1}{3} \cdot n + 1$, it is easy to conclude the following upper-bound for $\left|W_n^{\text{int}}\right|$.

Corollary 9.8. *For the family $\mathcal{G}_n$ it holds that*

$$\left|W_n^{\text{int}}\right| \leqslant |W[\mathcal{G}_n]| \leqslant \frac{1}{6} \cdot n^3 - \frac{1}{2} \cdot n^2 + \frac{1}{3} \cdot n + 1.$$

In lemmas that follow, we proceed by defining some intervals that are fully contained in $W[\mathcal{G}_n]$. We then try to tile these intervals so that their union form a bigger interval. The core part of estimating the length of W_n^{int} is the following lemma.

Lemma 9.9. *Fix positive integers $n \geqslant 7$ and $a_1, \ldots, a_k$ such that $P(a_1, a_2, \ldots, a_k)$ has n vertices and $a_1 + a_2 \geqslant 2\sqrt{n}$. Then*

$$[W(P(a_1, a_2, \ldots, a_k)), W(P(a_1+1, a_2-1, \ldots, a_k))] \subseteq W[\mathcal{G}_n].$$

Proof. For easier notation denote

$$\begin{aligned} G &= P(a_1, a_2, \ldots, a_k), \\ H &= P(a_1+1, a_2-1, \ldots, a_k) \end{aligned}$$

and observe that

$$W(H) - W(G) = n - 2a_1 - 3.$$

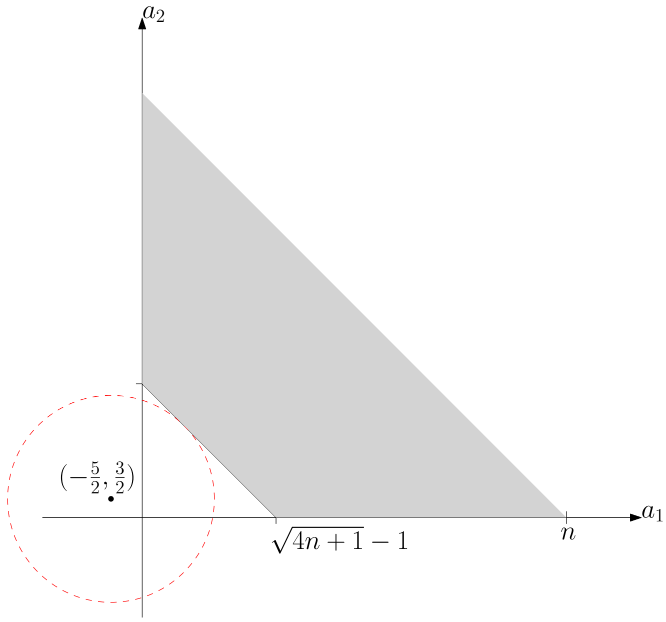


Figure 9.1: A graphic representation of the inequalities $a_1 + a_2 > \sqrt{1+4n} - 1$ (grayed surface) and $\left(a_1 + \frac{5}{2}\right)^2 + \left(a_2 - \frac{3}{2}\right)^2 > \frac{1}{2} + 2n$ (dashed circle). Notice that $a_1 + a_2 \leqslant n$.

Depending on which of both graphs H and G had smaller Wiener index we consider the following two cases. In both cases, we fill the space between $W(H)$ and $W(G)$ by adding at least $|n - 2a_1 - 3|$ additional edges to one of the graphs G or H that have bigger Wiener index, as described in Observation 9.6.

Case 1: $a_1 \geqslant \frac{n}{2} - \frac{3}{2}$. It holds that $W(H) \leqslant W(G)$, hence we fill the space between $W(H)$ and $W(G)$ by adding at least $|n - 2a_1 - 3|$ additional edges to G. It is therefore enough to show that $\binom{a_1}{2} \geqslant -(n - 2a_1 - 3)$. Indeed, since $n \geqslant 7$, we have

$$\binom{a_1}{2} + n - 2a_1 - 3 \quad = \quad \frac{\left(a_1 - \frac{5}{2}\right)^2}{2} - \frac{49}{8} + n \geqslant n - \frac{49}{8} > 0.$$

Case 2: $a_1 < \frac{n}{2} - \frac{3}{2}$. Now notice $W(H) > W(G)$, hence we fill the space between $W(H)$ and $W(G)$ by adding at least $|n - 2a_1 - 3|$ additional edges to H. From the fact that $a_1 + a_2 \geqslant 2\sqrt{n} > \sqrt{1+4n} - 1$, it is clear that

$$\left(a_1 + \frac{5}{2}\right)^2 + \left(a_2 - \frac{3}{2}\right)^2 > \frac{1}{2} + 2n. \tag{9.1}$$

$$\begin{aligned} D(n,i) &\longrightarrow P(1, n-i-2, 0, \ldots, 0) \\ &\longrightarrow P(2, n-i-3, 0, \ldots, 0) \\ &\ \vdots \\ &\longrightarrow D(n, i+1) \end{aligned}$$

Figure 9.2: The schema that describes the key graphs in the process of generating an interval W_n^{int}. The gap between any two graphs on the schema is filled by adding additional edges, as described in Observation 9.6. Note that all graphs on the figure are members of $\mathcal{G}_n$.

The same can be also observed on Figure 9.1, where the inequality (9.1) is drawn. Again, it is enough to show that $\binom{a_1+1}{2} + \binom{a_2-1}{2} \geqslant n - 2a_1 - 3$. By inequality (9.1) and since $a_1 \geqslant 0$, $n \geqslant 8$, we conclude

$$\begin{aligned} &\binom{a_1+1}{2} + \binom{a_2-1}{2} - (n - 2a_1 - 3) \\ = &\frac{\left(a_1 + \frac{5}{2}\right)^2 + \left(a_2 - \frac{3}{2}\right)^2}{2} - \frac{1}{4} - n \geqslant 0. \end{aligned}$$

□

With the tools provided, we now state the main result, where we iteratively transform our graph by a scheme, provided on Figure 9.2. Note that on each step, coming from graph G to G', Lemma 9.9 guarantees that all integers from $[W(G), W(G')]$ are members of $W[\mathcal{G}_n]$. We will use this schema to create a collection of intersecting intervals from $W[\mathcal{G}_n]$, which will be the building-blocks of W_n^{int}. Recall the Lemma 9.7 and define a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that

$$\begin{aligned} f(n,b) &= \binom{b+1}{3} + \left(\binom{b+1}{2} - 1\right) \cdot (n-b) + (n-b)^2 \\ &= \frac{1}{2}b^2 n - \frac{1}{3}b^3 + \frac{1}{2}b^2 - \frac{3}{2}bn + \frac{5}{6}b + n^2 - n. \end{aligned}$$

Now, consider the following bound.

Lemma 9.10. *It holds that*

$$W\left(D\left(n, \lfloor n - 2\sqrt{n} - 1 \rfloor\right)\right) \geqslant \frac{1}{6}n^3 - 2n^2 - \frac{1}{3}n^{\frac{3}{2}} + \frac{35}{6}n + \frac{7}{3}\sqrt{n}.$$

Proof. Let $p \in (0,1]$ be a real number, such that $\lfloor n - 2\sqrt{n} - 1 \rfloor = n - 2\sqrt{n} - 2 + p$. By substituting b with $n - 2\sqrt{n} - 2 + p$ in $f(n,b)$, we get

$$\begin{aligned} f\left(n, n-2\sqrt{n}-2+p\right) &= \frac{1}{6}n^3 - 2n^2 + \left(2p - \frac{1}{3}\right)n^{\frac{3}{2}} + \left(\frac{53}{6} - \frac{1}{2}p^2 - \frac{5}{2}p\right)n \\ &\quad + \left(\frac{31}{3} + 2p^2 - 10p\right)\sqrt{n} + 3 - \frac{1}{3}p^3 + \frac{5}{2}p^2 - \frac{31}{6}p \\ &\geqslant \frac{1}{6}n^3 - 2n^2 - \frac{1}{3}n^{\frac{3}{2}} + \frac{35}{6}n + \frac{7}{3}\sqrt{n}, \end{aligned}$$

as claimed. □

We are now ready to state the main result.

Theorem 9.11. *Let $W_n^{\text{int}} = [a,b]$ be a largest interval of contiguous integers such that $W_n^{\text{int}} \subseteq W[\mathcal{G}_n]$. Then, it holds*

$$a = \binom{n}{2} \quad \text{and} \quad b \geqslant \frac{1}{6}n^3 - \frac{5}{2}n^2 + O\left(n^{3/2}\right).$$

In particular, $\left|W_n^{\text{int}}\right| = \frac{1}{6}n^3 + O\left(n^2\right)$.

Proof. We prove the theorem by using Lemma 9.9, which provides us with a collection of intervals $[W(D(n,i)), W(D(n,i+1))]$ for $i \in \left[1, \lfloor n - 2\sqrt{n} - 1 \rfloor\right]$, see Figure 9.2.

First, consider the left end of the interval with $i = 1$. By iteratively adding the edges, we can easily extend the left end from $W(D(n,1)) = (n-1)^2$ to $W(K_n) = \binom{n}{2}$. Since by Claim 9.4, $\binom{n}{2}$ is the minimum of $W[\mathcal{G}_n]$, we cannot improve the lower-bound of this interval any further.

Now, consider the right part of the interval and calculate the lower bound of $W(D(n,k))$ with $k = \left\lfloor n - \sqrt{2n} - 1 \right\rfloor$. Using Lemmas 9.7 and 9.10, we have

$$W\left(D\left(n, \lfloor n - 2\sqrt{n} - 1 \rfloor\right)\right) \geqslant \frac{1}{6}n^3 - 2n^2 - \frac{1}{3}n^{\frac{3}{2}} + \frac{35}{6}n + \frac{7}{3}\sqrt{n}.$$

From this, we subtract the left-end of the interval, and obtain a lower-bound of the cardinality of W_n^{int}, i.e.

$$\begin{aligned} \left|W_n^{\text{int}}\right| &\geqslant \frac{1}{6}n^3 - 2n^2 - \frac{1}{3}n^{\frac{3}{2}} + \frac{35}{6}n + \frac{7}{3}\sqrt{n} - \binom{n}{2} \\ &= \frac{1}{6}n^3 - \frac{5}{2}n^2 - \frac{1}{3}n^{\frac{3}{2}} + \frac{19}{3}n + \frac{7}{3}\sqrt{n}, \end{aligned} \tag{9.2}$$

which concludes the proof of our theorem. □

The following corollary clearly follows.

Corollary 9.12. *Let $W[\mathcal{G}_n]$ be an image of a Wiener index function on the set $\mathcal{G}_n$. Then*

$$|W[\mathcal{G}_n]| = \frac{1}{6}n^3 + O\left(n^2\right).$$

Let us also note that whenever $n \geqslant 25$ the lower bound from (9.2) is always always greater than a half of our trivial upper-bound form Corollary 9.8, which also implies uniqueness of the observed interval for $n \geqslant 25$.

9.3 CONCLUDING REMARKS AND FURTHER WORK

From Example 9.5, we observed that $W[\mathcal{G}_4] = W_4^{\text{int}} = \left[\binom{4}{2}, \binom{5}{3}\right]$. Similarly, one can easily check that also for other values of $n \leqslant 4$ it holds that $W[\mathcal{G}_n] = W_n^{\text{int}} = \left[\binom{n}{2}, \binom{n+1}{3}\right]$. When estimating the cardinality of W_n^{int}, one could improve the final result from (9.2) by precisely calculating the gap that we made with an inequality from (9.1). Also, one could improve the cardinality of $W[\mathcal{G}_n]$ by addressing the fact that the intervals of type $\left[W(D(n,i)) - \binom{n-i}{2}, W(D(n,i))\right] \in W[\mathcal{G}_n]$ are disjoint inside the area of red-dashed circle from Figure 9.1. Summing these would increase the lower bound of cardinality of $W[\mathcal{G}_n]$. One implication of these optimizations may be the proof of uniqueness also for $n \leqslant 24$.

Conjecture 9.13. *Let n be a positive integer. Interval W_n^{int} is unique and starts at $\binom{n}{2}$.*

The complementary question we are interested in is also the cardinality of $\left[\binom{n}{2}, \binom{n+1}{3}\right] \setminus W[\mathcal{G}_n]$. Notice, that the cardinality of $\left[\binom{n}{2}, \binom{n+1}{3}\right] \setminus W[\mathcal{G}_n]$ is at least linear, since the gap between the two graphs in $\mathcal{G}_n$ with highest Wiener index equals $n-4$ (for $n \geqslant 4$). We believe that the number of values $W[\mathcal{G}_n]$ misses is indeed linear, hence the following conjecture.

Conjecture 9.14. *The cardinality of $W[\mathcal{G}_n]$ is of order $\frac{1}{6}n^3 - \frac{1}{2}n^2 + \Theta(n)$.*

Among other generalizations one could also answer similar questions on the family of all trees on n vertices. For a fixed n, we define $W[\mathcal{T}_n]$ to be the image of W under $\mathcal{T}_n$, i.e.

$$i \in W[\mathcal{T}_n] \Leftrightarrow \text{ there is a graph } G \in \mathcal{T}_n \text{ such that } W(G) = i.$$

We conjecture the following.

Conjecture 9.15. *The cardinality of $W[\mathcal{T}_n]$ equals $\frac{1}{6}n^3 + \Theta\left(n^2\right)$.*

Conjecture 9.16. *In the family of $\mathcal{T}_n$, the cardinality of W_n^{int} equals $\Theta\left(n^3\right)$.*

10

FUTURE WORK

In this chapter, we point out some of interesting open problems and briefly present additional work in progress on betweenness centralization. We further discuss some of the results from previous chapters, and summarize conjectures on the extremal graphs for Freeman centralization.

10.1 MAX-DEGREE PROPERTY

An interesting question for studying centrality indices is presented in [48], where authors are asking the following.

Problem 10.1. In the class of graphs with fixed maximum degree $\Delta(G)$ and fixed number of edges, is it true that for any extremal-pair (G, v) for betweenness centralization it holds $\deg(v) = \Delta(G)$?

While the question above will be further discussed later, one may similarly inspect if a vertex from an extremal pair for some other family of graph (and other centrality index) is of highest degree. Furthermore, for group centrality measures one can ask if all maximizing sets contain a vertex of maximum degree. As presented in the thesis, the mentioned property holds

- for most centralities (betweenness, degree, closeness, eccentricity) for the family $\mathcal{G}_n$,
- for most centralizations (betweenness, degree, closeness, eccentricity) for the family $\mathcal{G}_n$,
- for most centralizations (betweenness, degree, closeness, eccentricity) for the family $\mathscr{B}(n_0, n_1)$,
- for degree group centralization for the family $\mathcal{G}_n$, on any size of the set,
- for eccentricity group centralization on 2-sets, for the family $\mathcal{G}_n$.

In particular, we know that in $\mathcal{G}_n$, for any extremal pair (G, v) for betweenness centralization, it holds $\deg_G(v) = \Delta(G)$. However, this does not answer the question above. Let us briefly present the work in progress, that partly gives answer to the Problem 10.1.

10.2 BETWEENNESS CENTRALIZATION

Let us recall that $B_1(v, G)$ stands for betweenness centralization, see Definition 3.11 on page 29. In real social networks, nodes often have a (natural or enforced) maximum number of connections. It is, therefore, interesting to consider the problem of maximizing betweenness centralization in classes of graphs where we impose an upper bound on the maximum degree. Let

$$\mathcal{H}(\Delta, m) := \{G : |E(G)| = m \text{ and } 1 \leqslant \deg(v) \leqslant \Delta \text{ for every } v \in V(G)\},$$

be the class of graphs on exactly m edges and maximum degree at most Δ that contain no isolated vertices. Observe that the extra condition regarding isolated vertices is necessary as one can artificially inflate the betweenness centralization index by adding dummy nodes.

The following has been conjectured in [48]:

Conjecture 10.2. *For fixed integers m and Δ, let (G^*, v^*) be the optimizers of*

$$\max_{G \in \mathcal{H}} \max_{v \in G} B_1(v; G).$$

Then $\deg_{G^*}(v^*) = \Delta$.

Conjecture 10.3. *Let $H^*_{\Delta,m}$ be the minimizer of $\min_{G \in \mathcal{H}(\Delta,m)} W(G)$ and let $(G^*_{\Delta,m}, v^*)$ be the optimizers of*

$$\max_{G \in \mathcal{H}(\Delta,m)} \max_{v \in V(G)} B(v; G).$$

Then $G^*_{\Delta,m}$ can be constructed by taking S_Δ and identifying all leaves with any vertex of graph $H^*_{\Delta, \frac{m-\Delta}{\Delta}}$.

We now briefly outline the work in progress and the partial results related to Conjecture 10.2.

We can prove that if $m \geqslant n \log n$, the conjecture holds. We do this by finding an upper bound on $B(v)$ for a vertex of degree at most $\Delta - 1$ and exhibiting a graph (actually, a tree) which contains a vertex that exceeds this bound.

The upper bound is attained by using Turán's theorem [133] as an upper bound on the cutting number of a vertex. Using some known bounds on Turán's numbers $t_k(n)$, this results in the proof of the following lemma.

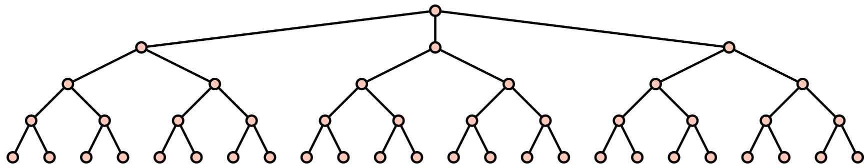


Figure 10.1: An example of a tree T^* with $n = 46$ and $\Delta = 3$.

Lemma 10.4. *Let G be a graph on n vertices and let $v \in V(G)$ be a vertex of degree k. Then,*

$$B(v; G) \leqslant t_k(n-1) \leqslant \frac{(n-1)^2}{2}\left(1 - \frac{1}{k}\right).$$

In terms of Freeman centralization, we therefore have the following bound.

Corollary 10.5. *Let $G \in \mathcal{H}(\Delta, m)$ and let $v \in G$ with $\deg_G(v) = k < \Delta$. Then*

$$B_1(v; G) \leqslant n \cdot B(v; G) < \frac{n^3}{2}\left(1 - \frac{1}{k}\right) \leqslant \frac{n^3}{2}\left(1 - \frac{1}{\Delta - 1}\right).$$

Next, we construct a rooted tree, T^*, and estimate the betweenness centralization index of its root to provide a lower bound on the value of betweenness centralization. The tree T^* is a balanced tree with almost all vertices of degree either Δ or 1, see example on Figure 10.1, with $n = 46$ and $\Delta = 3$. We then show that, in most cases (unless $m - n$ is "small"), this constructive lower bound is greater than the above obtained upper bound. Specifically, with help of the tree from Figure 10.1, we prove the following lemma.

Lemma 10.6. *It holds that*

$$\max_{G \in \mathcal{H}(\Delta,m)} \max_{v \in V(G)} B_1(v; G) \geqslant \frac{m^3}{2}\left(1 - \frac{1}{\Delta}\right) - m^2 \log m.$$

If (G^*, v^*) is an optimizer for $\max_{G \in \mathcal{H}(\Delta,m)} \max_{v \in V(G)} B_1(v; G)$, and supposing that $\deg(v^*) < \Delta$, then it clearly follows

$$\begin{aligned} \frac{n^3}{2}\left(1 - \frac{1}{\Delta - 1}\right) &\geqslant B_1(v^*, G^*) \geqslant \frac{m^3}{2}\left(1 - \frac{1}{\Delta}\right) - m^2 \log m \\ n^3 &> m^3 - 2m^2 \log m. \end{aligned} \tag{10.1}$$

If $\Delta \leqslant 3$, or if $n < 6$, this is not strictly true. However, for these cases the conjecture can be easily checked by hand. It follows that (10.1) gives a desired contradiction in most of the cases, in particular, whenever $m > n \log n$.

In order to prove the conjecture, one need to take care of the cases where $n \approx m$, or improve the bounds from the above lemmas.

In next section we discuss some of the other conjectures from Chapters 4 on page 35– 7 on page 89.

	THE FAMILY $\mathcal{G}_n$	THE FAMILY $\mathscr{B}(n_0, n_1)$
Degree	Freeman [1979]	Borgatti and Everett [1997]
Closeness	Everett et al. [2004]	Theorem 4.3 on page 40
Eccentricity	Proposition 5.5 on page 60	Theorem 5.3 on page 58
Betweenness	Freeman [1979]	Sinclair [2005]
Transmission	Theorem 6.8 on page 81	?

Table 10.1: The list of results on extremal network for Freeman centralization of some centrality indices. Note that Theorem 5.3 only describe the *acyclic* extremal graphs for eccentricity centralization.

10.3 CONCLUDING REMARKS

The reader can notice that most of our results are focused towards Freeman centralization of some well-known centrality indices, such as degree, betweenness, eccentricity, closeness and transmission. For these indices, we tried to determine networks that maximize or in some cases minimize the Freeman centralization. For the family of $\mathcal{G}_n$ (all connected networks on n nodes) and the family of $\mathscr{B}(n_0, n_1)$ (all two-mode networks on fixed bipartition sizes n_0 and n_1) some results on extremal networks are summarized in Table 10.1 (note that $\mathcal{G}_n$ represents the family of all connected networks on n vertices, while $\mathscr{B}$ is the family of all connected bipartite graphs with fixed bipartition sizes).

In addition to the results above, in thesis we also describe

- (some) networks that maximize the value of eccentricity centralization for networks with fixed maximum degree and fixed number of vertices,
- networks that maximize the value of eccentricity centralization for networks with fixed number of edges,
- networks that minimize the value of transmission centralization for $\mathcal{G}_n$,
- (some) networks that maximize the value of group eccentricity centralization (for groups of size two), for $\mathcal{G}_n$,
- networks that maximize the value of group betweenness centralization (for groups of size two), for $\mathcal{G}_n$, and
- networks that maximize the value of group degree centralization in $\mathcal{G}_n$ (for arbitrary size of the group).

We are aware that our analysis for extremal graphs in Freeman centralization is just a small portion of centrality research area. One of the aims of our research

CENTRALITY INDEX	CONJECTURED RESULTING EXTREMAL GRAPH	CLASS OF GRAPHS	REF.
Group betweenness centralization with groups of sizes k.	$S_{n-k,k}$ (definition on page 100)	$\mathcal{G}_n$	[84]
Group eccentricity centralization with groups of sizes $k \leqslant \frac{n}{3}$.	$D(n, 3c-1)$ (definition on page 10)	$\mathcal{G}_n$	[84]
Eigenvalue centralization (see Section 3.4.6 on page 32)	$H(\cdot, n_0, n_1)$ (definition on page 11)	$\mathscr{B}(n_0, n_1)$	[48]
Betweenness centralization	$G^*_{\Delta,m}$ (definition on page 130)	$\mathcal{H}(\Delta, m)$	[48]

Table 10.2: A partial list of open conjectures on the extremal graphs for Freeman centralization.

is to provide some additional insight on the fundamental differences between various types of centrality. Let us also note that there remain quite some open problems in this specific field of characterizing extremal graphs. Some of these are summarized on Table 10.2.

APPENDIX

A

BIBLIOGRAPHY

[1] James Abello, Adam L. Buchsbaum, and Jeffery R. Westbrook. A functional approach to external graph algorithms. In *Algorithms–ESA'98*, pages 332–343. Springer, 1998. (Cited on page 14.)

[2] Woo-kyoung Ahn, Nancy S. Kim, Mary E. Lassaline, and Martin J. Dennis. Causal status as a determinant of feature centrality. *Cognitive Psychology*, 41(4):361–416, 2000. (Cited on pages 1 and 21.)

[3] Sebastian E. Ahnert and Thomas Fink. Clustering signatures classify directed networks. *Physical Review E*, 78(3):036112, 2008. (Cited on page 16.)

[4] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. (Cited on pages 16, 20, and 21.)

[5] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999. (Cited on page 16.)

[6] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. (Cited on page 18.)

[7] Jac M. Anthonisse. The rush in a directed graph. Technical report, Stichting Mathematisch Centrum, Amsterdam, October 1971. URL `http://oai.cwi.nl/oai/asset/9791/9791A.pdf`. (Cited on page 28.)

[8] Mustapha Aouchiche and Pierre Hansen. Proximity and remoteness in graphs: results and conjectures. *Networks*, 58(2):95–102, 2011. (Cited on page 55.)

[9] Kamal Badar, Julie M. Hite, and Yuosre F. Badir. Examining the relationship of co-authorship network centrality and gender on academic research performance: the case of chemistry researchers in pakistan. *Scientometrics*, 94(2):755–775, 2013. (Cited on pages 1 and 22.)

[10] Yih-En Andrew Ban, Sergei Bespamyatnikh, and Nabil H. Mustafa. On a conjecture on Wiener indices in combinatorial chemistry. In *Computing and combinatorics*, pages 509–518. Springer, 2003. (Cited on page 121.)

[11] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. (Cited on pages 14 and 15.)

[12] Barbara Bauer, Ferenc Jordán, and János Podani. Node centrality indices in food webs: Rank orders versus distributions. *Ecological Complexity*, 7 (4):471–477, 2010. (Cited on pages 1 and 22.)

[13] Alex Bavelas. A mathematical model for group structures. *Hum. Organ.*, 7(3):16–30, 1948. (Cited on page 22.)

[14] Alex Bavelas. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.*, 22(6):725–730, 1950. (Cited on page 35.)

[15] Alex Bavelas. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.*, pages 725–730, 1950. (Cited on pages 22 and 29.)

[16] Murray Beauchamp. An improved index of centrality. *Behavioral Science*, 10(2):161–163, 1965. (Cited on page 35.)

[17] Jocelyn R. Bell. Subgroup centrality measures. *Network Science*, 2(02): 277–297, 2014. (Cited on pages 24, 90, and 152.)

[18] Zhiqiang Bi, Christos Faloutsos, and Flip Korn. The dgx distribution for mining massive, skewed data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26. ACM, 2001. (Cited on page 15.)

[19] Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004. (Cited on page 16.)

[20] Stephen P Borgatti. Identifying sets of key players in a social network. *Comput. Math. Organ. Theory*, 12(1):21–34, 2006. (Cited on pages 23 and 90.)

[21] Stephen P. Borgatti and Martin G. Everett. Network analysis of 2-mode data. *Soc. Netw.*, 19:243–269, 1997. (Cited on pages 9, 132, and 152.)

[22] Rodrigo A. Botafogo, Ehud Rivlin, and Ben Shneiderman. Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180, 1992. (Cited on page 35.)

[23] Ulrik Brandes. A faster algorithm for betweenness centrality*. *J. Math. Sociol.*, 25(2):163–177, 2001. (Cited on page 89.)

[24] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Soc. Netw.*, 30(2):136–145, 2008. (Cited on page 89.)

[25] Ulrik Brandes and Thomas Erlebach. *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media, 2005. (Cited on page 22.)

[26] Ulrik Brandes, Patrick Kenis, and Dorothea Wagner. Centrality in policy network drawings. In *Graph Drawing*, pages 250–258. Springer, 1999. (Cited on page 26.)

[27] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000. (Cited on pages 14 and 16.)

[28] Carter T. Butts. Exact bounds for degree centralization. *Soc. Netw.*, 28(4): 283–296, 2006. (Cited on pages 25, 26, 56, 87, and 152.)

[29] Gilles Caporossi, Marcia Paiva, Damir Vukičević, and Marcelo Segatto. Centrality and betweenness: vertex and edge decomposition of the Wiener index. *MATCH Commun. Math. Comput. Chem.*, 68(1):293–302, 2012. (Cited on page 75.)

[30] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-mat: A recursive model for graph mining. In *SDM*, volume 4, pages 442–446. SIAM, 2004. (Cited on page 15.)

[31] Fan Chung and Linyuan Lu. The diameter of sparse random graphs. *Adv. Appl. Math.*, 26(4):257–279, 2001. (Cited on page 20.)

[32] Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. U.S.A.*, 99(25):15879–15882, 2002. (Cited on page 16.)

[33] Eduardo Cotilla-Sanchez, Paul D. H. Hines, Clayton Barrows, and Seth Blumsack. Comparing the topological and electrical structure of the north american electric power infrastructure. *Systems Journal, IEEE*, 6 (4):616–626, 2012. (Cited on pages 1 and 22.)

[34] Carlos M da Fonseca, Mohammad Ghebleh, Ali Kanso, and Dragan Stevanović. Counterexamples to a conjecture on Wiener index of common

neighborhood graphs. *MATCH Commun. Math. Comput. Chem.*, 72:333–338, 2014. (Cited on page 31.)

[35] Peter Dankelmann and Simon Mukwembi. Upper bounds on the average eccentricity. *Discrete Appl. Math.*, 167:72–79, 2014. (Cited on page 55.)

[36] Peter Dankelmann, David Erwin, Wayne Goddard, Simon Mukwembi, and Henda C. Swart. Eccentric counts, connectivity and chordality. *Inform. Process. Lett*, 112(24):944–947, 2012. (Cited on page 55.)

[37] King David. *The Holy Bible English Standard Edition (ESV)*, chapter Psalms. Crossway Bibles, a publishing ministry of Good News Publishers., 2002. Ps 90, 12. (Cited on page v.)

[38] Allison Davis, Burleigh Bradford Gardner, and Mary R. Gardner. *Deep south*. University of Chicago Press, 1969. (Cited on pages 14 and 36.)

[39] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001. (Cited on page 16.)

[40] Shlomi Dolev, Yuval Elovici, Rami Puzis, and Polina Zilberman. Incremental deployment of network monitors based on group betweenness centrality. *Inform. Process. Lett.*, 109(20):1172–1176, 2009. (Cited on page 89.)

[41] Sergey N. Dorogovtsev, A. V. Goltsev, and José Ferreira F. Mendes. Pseudofractal scale-free web. *Physical Review E*, 65(6):066122, 2002. (Cited on page 18.)

[42] Zhibin Du and Aleksandar Ilić. On AGX conjectures regarding average eccentricity. *MATCH Commun. Math. Comput. Chem.*, 69:579–587, 2013. (Cited on page 55.)

[43] Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, 34(2):33–40, 2005. (Cited on pages 13 and 151.)

[44] Paul Erdős and Alfréd Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959. (Cited on page 19.)

[45] Ernesto Estrada and Juan A. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005. (Cited on pages 23 and 90.)

[46] Martin G. Everett and Stephen P. Borgatti. The centrality of groups and classes. *J. Math. Sociol.*, 23(3):181–201, 1999. (Cited on pages ix, 2, 23, 24, 26, 30, 90, 92, 100, 107, 151, 152, and 156.)

[47] Martin G. Everett and Stephen P. Borgatti. Extending centrality. *Models and methods in social network analysis*, 35(1):57–76, 2005. (Cited on pages 23 and 90.)

[48] Martin G. Everett, Philip Sinclair, and Peter Dankelmann. Some centrality results new and old. *J. Math. Sociol.*, 28(4):215–227, 2004. (Cited on pages vii, ix, 2, 11, 25, 35, 36, 39, 40, 52, 58, 60, 87, 104, 129, 130, 132, 133, 152, and 153.)

[49] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999. (Cited on pages 14 and 15.)

[50] Jiří Fink, Borut Lužar, and Riste Škrekovski. Some remarks on inverse Wiener index problem. *Discrete Appl. Math.*, 160(12):1851–1858, 2012. (Cited on page 121.)

[51] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977. (Cited on pages 28 and 55.)

[52] Linton C. Freeman. Centrality in social networks conceptual clarification. *Soc. Netw.*, 1(3):215–239, 1979. (Cited on pages vii, ix, 24, 25, 27, 29, 35, 56, 75, 87, 89, 91, 100, 108, 132, 151, and 152.)

[53] Linton C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press (July 23, 2004), 2004. (Cited on page 1.)

[54] Zoltán Füredi and Younjin Kim. The structure of the typical graphs of given diameter. *Discrete Math.*, 313(2):155–163, 2013. (Cited on page 55.)

[55] Xianya Geng, Shuchao Li, and Meng Zhang. Extremal values on the eccentric distance sum of trees. *Discrete Appl. Math.*, 161(16):2427–2439, 2013. (Cited on page 55.)

[56] Edgar N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144, 12 1959. doi: 10.1214/aoms/1177706098. (Cited on page 19.)

[57] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826, 2002. (Cited on page 18.)

[58] Daniel Gomez, Enrique González-Arangüena, Conrado Manuel, Guillermo Owen, Monica del Pozo, and Juan Tejada. Centrality and power in social networks: a game theoretic approach. *Math. Soc. Sci.*, 46 (1):27–54, 2003. (Cited on pages 1 and 21.)

[59] Roger Guimera, Stefano Mossa, Adrian Turtschi, and L. A. Nunes Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(22):7794–7799, 2005. (Cited on pages 1 and 21.)

[60] Ivan Gutman and Tamas Kortvelyesi. Wiener indices and molecular surfaces. *Z. Naturforsch. A.*, 50(7):669–671, 1995. (Cited on page 31.)

[61] Ivan Gutman and Yeong-Nan Yeh. The sum of all distances in bipartite graphs. *Math. Slovaca*, 45(4):327–334, 1995. (Cited on page 121.)

[62] Per Hage and Frank Harary. *Structural models in anthropology*. Cambridge University Press, 1983. (Cited on page 1.)

[63] Per Hage and Frank Harary. Eccentricity and centrality in networks. *Soc. Netw.*, 17(1):57–63, 1995. doi: http://dx.doi.org/10.1016/0378-8733(94)00248-9. (Cited on pages 27, 56, and 93.)

[64] Seifollah L. Hakimi and S. N. Maheshwari. Optimum locations of centers in networks. *Operations Res.*, 20:967–973, 1972. ISSN 0030-364X. (Cited on page 26.)

[65] Bronwyn H. Hall, Adam B. Jaffe, and Manuel Trajtenberg. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. NBER Working Papers 8498, National Bureau of Economic Research, Inc, October 2001. URL http://ideas.repec.org/p/nbr/nberwo/8498.html. (Cited on page 117.)

[66] Paul W. Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comp. Group Stud.*, 1971. (Cited on page 16.)

[67] Katarına Hrinakova, Martin Knor, Riste Škrekovski, and Aleksandra Tepeh. A congruence relation for the Wiener index of graphs with a tree-like structure. *MATCH Commun. Math. Comput. Chem.*, 72:791–806, 2014. (Cited on page 31.)

[68] Hongbo Hua, Shenggui Zhang, and Kexiang Xu. Further results on the eccentric distance sum. *Discrete Appl. Math.*, 160(1):170–180, 2012. (Cited on page 55.)

[69] Bernardo A. Huberman and Lada A. Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999. (Cited on page 14.)

[70] Aleksandar Ilić. On the extremal properties of the average eccentricity. *Comput. Math. Appl.*, 64(9):2877–2885, 2012. (Cited on page 55.)

[71] Aleksandar Ilić, Guihai Yu, and Lihua Feng. On the eccentric distance sum of graphs. *J. Math. Anal. Appl.*, 381(2):590–600, 2011. (Cited on page 55.)

[72] Matthew O. Jackson. *Social and economic networks*, volume 3. Princeton University Press Princeton, 2008. (Cited on pages 1 and 21.)

[73] Hawoong Jeong, Sean P Mason, Albert-László Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833): 41–42, 2001. (Cited on pages 1 and 21.)

[74] Camille Jordan. Sur les assemblages de lignes. *J. Reine Angew. Math*, 70 (185):81, 1869. (Cited on pages 1 and 21.)

[75] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. (Cited on page 32.)

[76] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999. (Cited on page 14.)

[77] Martin Knor and Riste Škrekovski. Wiener index of generalized 4-stars and of their quadratic line graphs. *Australas. J. Comb.*, 58:119–126, 2014. (Cited on page 31.)

[78] Martin Knor, Borut Lužar, Riste Škrekovski, and Ivan Gutman. On Wiener index of common neighborhood graphs. *MATCH Commun. Math. Comput. Chem.*, 72:321–332, 2014. (Cited on page 31.)

[79] Martin Knor, Martin Mačaj, Primož Potočnik, and Riste Škrekovski. Complete solution of equation $W(L^3(T)) = W(T)$ for the Wiener index of iterated line graphs of trees. *Discrete Appl. Math.*, 171:90–103, 2014. (Cited on page 31.)

[80] Martin Knor, Martin Mačaj, Primož Potočnik, and Riste Škrekovski. Trees T satisfying $W(L^3(T) = W(T)$. *Filomat*, 28(3):551–556, 2014. (Cited on page 31.)

[81] Martin Knor, Aleksandra Tepeh, and Riste Škrekovski. Mathematical aspects of Wiener index. *To appear in Ars Math. Contemp.*, 2015. (Cited on page 31.)

[82] Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In *Network analysis*, pages 16–61. Springer, 2005. (Cited on page 22.)

[83] Matjaž Krnc and Riste Škrekovski. Algorithmic approach to degree centralization in large networks. *Submitted for publication*, 2015. (Cited on page xi.)

[84] Matjaž Krnc and Riste Škrekovski. Group centralization of network indices. *Discrete Appl. Math.*, 186:147–157, 2015. (Cited on pages xi, 28, 107, 108, and 133.)

[85] Matjaž Krnc and Riste Škrekovski. Centralization of transmission in networks. *Discrete Math.*, 338(12):2412–2420, 2015. (Cited on page xi.)

[86] Matjaž Krnc and Riste Škrekovski. On the Wiener inverse interval problem. *Submitted for publication*, 2015. (Cited on page xi.)

[87] Matjaž Krnc, Jean-Sébastien Sereni, Riste Škrekovski, and Zelealem Yilma. Maximizing betweenness centralization in graphs with bounded maximum degree. *In preparation*, 2015. (Cited on page xi.)

[88] Matjaž Krnc, Jean-Sébastien Sereni, Riste Škrekovski, and Zelealem Yilma. Centralization measures for two-mode data of prescribed sizes. *Submitted for publication*, 2015. (Cited on page xi.)

[89] Matjaž Krnc, Jean-Sébastien Sereni, Riste Škrekovski, and Zelealem Yilma. Eccentricity of networks with structural constraints. *Submitted for publication*, 2015. (Cited on page xi.)

[90] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer netw.*, 31(11):1481–1493, 1999. (Cited on page 14.)

[91] Mirko Lepović and Ivan Gutman. A collective property of trees and chemical trees. *J. Chem. Inf. Comput. Sci.*, 38(5):823–826, 1998. (Cited on page 121.)

[92] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014. (Cited on pages 116 and 117.)

[93] Jure Leskovec and Julian J. Mcauley. Learning to discover social circles in ego networks. In P. Bartlett, F.c.n. Pereira, C.j.c. Burges, L. Bottou, and K.q. Weinberger, editors, *Advances in Neural Information Processing*

Systems 25, pages 548–556. 2012. URL http://books.nips.cc/papers/files/nips25/NIPS2012_0272.pdf. (Cited on page 117.)

[94] Jure Leskovec and Rok Sosič. SNAP: A general purpose network analysis and graph mining library in C++. http://snap.stanford.edu/snap, June 2014. (Cited on page 117.)

[95] Jure Leskovec and Rok Sosič. Snap.py: SNAP for Python, a general purpose network analysis and graph mining tool in Python. http://snap.stanford.edu/snappy, June 2014. (Cited on page 117.)

[96] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM. ISBN 1-59593-135-X. (Cited on page 117.)

[97] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007. ISSN 1559-1131. (Cited on page 117.)

[98] Pedro G Lind, Marta C González, and Hans J Herrmann. Cycles and clustering in bipartite networks. *Physical review E*, 72(5):056127, 2005. (Cited on page 16.)

[99] Mark Lorenzen and Kristina Vaarst Andersen. Centrality and creativity: does richard florida's creative class offer new insights into urban hierarchy? *Economic Geography*, 85(4):363–390, 2009. (Cited on pages 1 and 21.)

[100] Barrett Lyon. The web-graph. (Cited on pages 14, 16, and 17.)

[101] J. A. McHugh. *Algorithmic Graph Theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1990. ISBN 0-13-023615-2. (Cited on pages 27 and 55.)

[102] B. McKay. Finding maximum of a function with unfixed number of variables, 2014. URL http://mathoverflow.net/q/190447. (Cited on page 85.)

[103] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967. (Cited on pages 16 and 121.)

[104] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. (Cited on page 18.)

[105] Ron Milo, Nadav Kashtan, Shalev Itzkovitz, Mark E. J. Newman, and Uri Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv preprint cond-mat/0312028*, 2003. (Cited on pages 18 and 20.)

[106] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007. (Cited on page 117.)

[107] Eiji Miyano and Hirotaka Ono. Maximum domination problem. In *Proceedings of the Seventeenth Computing: The Australasian Theory Symposium-Volume 119*, pages 55–62. Australian Computer Society, Inc., 2011. (Cited on pages vii, x, 2, 92, 107, 111, 115, 116, 119, and 157.)

[108] Robert L. Moxley and Nancy F. Moxley. Determining point-centrality in uncontrived social networks. *Sociometry*, 37(1):122–130, 1974. (Cited on page 35.)

[109] Mark E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci.*, 101(suppl 1):5200–5205, 2004. (Cited on pages 13 and 151.)

[110] Mark E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. In *Complex networks*, pages 337–370. Springer, 2004. (Cited on pages 13 and 151.)

[111] Mark E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8577–8582, 2006. (Cited on page 18.)

[112] Mark E. J. Newman. *Networks*. Oxford University Press, Oxford, 2010. ISBN 978-0-19-920665-0. doi: 10.1093/acprof:oso/9780199206650.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780199206650.001.0001. An introduction. (Cited on page 55.)

[113] Mark E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036–122, 2003. (Cited on page 151.)

[114] U. J. Nieminen. On the centrality in a directed graph. *Soc. Sci. Res.*, 2(4):371–378, 1973. (Cited on page 35.)

[115] Jukka-Pekka Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and Albert-László Barabási.

Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104(18):7332–7336, 2007. (Cited on pages 1, 21, and 121.)

[116] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Soc. netw.*, 31(2):155–163, 2009. (Cited on page 16.)

[117] Oskar Perron. Zur theorie der matrices. *Math. Ann.*, 64(2):248–263, 1907. (Cited on page 32.)

[118] John J. Potterat, L. Phillips-Plummer, Stephen Q. Muth, R. B. Rothenberg, D. E. Woodhouse, T. S. Maldonado-Long, H. P. Zimmerman, and J. B. Muth. Risk network structure in the early epidemic phase of HIV transmission in colorado springs. *Sex. Transm. Infect.*, 78(suppl 1):i159–i163, 2002. (Cited on pages 13 and 151.)

[119] C. H. Proctor and C. P. Loomis. *Research Methods in Social Relations*, chapter Analysis of sociometric data, pages 561–586. Dryden Press, New York, 1951. (Cited on page 22.)

[120] Natasa Pržulj, Derek G. Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004. (Cited on page 18.)

[121] Rami Puzis, Yuval Elovici, and Shlomi Dolev. Finding the most prominent group in complex networks. *AI communications*, 20(4):287–296, 2007. (Cited on page 89.)

[122] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026–112, 2003. (Cited on page 18.)

[123] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and Albert-László Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002. (Cited on page 18.)

[124] Sidney Redner. How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. J. B*, 4(2):131–134, 1998. (Cited on page 14.)

[125] D. H. Rouvray and B. C. Crafford. The dependence of physical-chemical properties on topological factors. *S. Afr. J. Sci.*, 72(2):47–51, Feb 1976. (Cited on page 31.)

[126] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4): 581–603, 1966. (Cited on page 35.)

[127] John R. Seeley. The net of reciprocal influence; a problem in treating sociometric data. *Can. J. Psychol.*, (3):234–240, 1949. (Cited on page 22.)

[128] Philip Sinclair. Betweenness centralization for bipartite graphs. *J. Math. Sociol.*, 29(1):25–31, 2004. (Cited on pages 39, 58, and 152.)

[129] Philip Sinclair. Erratum: Betweenness centralization for bipartite graphs. *J. Math. Sociol.*, 29(3):263–264, 2005. ISSN 0022-250X. doi: 10.1080/00222500590961630. (Cited on pages 58, 132, and 152.)

[130] Riste Škrekovski and Ivan Gutman. Vertex version of the Wiener theorem. *MATCH Commun. Math. Comput. Chem.*, 72:295–300, 2015. (Cited on page 76.)

[131] Leonard I Stiel and George Thodos. The normal boiling points and critical constants of saturated aliphatic hydrocarbons. *The American Institute of Chemical Engineers*, 8(4):527–529, 1962. (Cited on page 31.)

[132] Sudhir Leslie Tauro, Christopher Palmer, Georgos Siganos, and Michalis Faloutsos. A simple conceptual model for the internet topology. In *Global Telecommunications Conference, 2001. GLOBECOM'01. IEEE*, volume 3, pages 1667–1671. IEEE, 2001. (Cited on page 15.)

[133] Paul Turán. On an extremal problem in graph theory. *Mat. Fiz. Lapok*, 48 (436-452):137, 1941. (Cited on page 130.)

[134] William Thomas Tutte. *Graph theory*, volume 21 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2001. Reprint of the 1984 original. (Cited on pages 36 and 152.)

[135] Thomas W. Valente and Robert K. Foreman. Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. *Soc. Netw.*, 20(1):89–105, 1998. (Cited on page 35.)

[136] Stephan G. Wagner. A class of trees and its Wiener index. *Acta Appl. Math.*, 91(2):119–132, 2006. (Cited on page 121.)

[137] Hua Wang and Guang Yu. All but 49 numbers are Wiener indices of trees. *Acta Appl. Math.*, 92(1):15–20, 2006. (Cited on page 121.)

[138] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994. (Cited on pages 27 and 55.)

[139] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998. (Cited on pages 16 and 21.)

[140] A. Weber. *Über den Standort der Industrien*. J. C. B. Mohr, Tübingen, 1909. (Cited on page 26.)

[141] Scott White and Padhraic Smyth. Algorithms for estimating relative importance in networks. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–275. ACM, 2003. (Cited on pages 23 and 90.)

[142] Harry Wiener. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, 69(1):17–20, 1947. (Cited on pages 30, 31, 76, and 121.)

[143] Wikipedia. Network motif — wikipedia, the free encyclopedia, 2015. [Online; accessed 26-May-2015]. (Cited on page 18.)

[144] Baoyindureng Wu. Wiener index of line graphs. *MATCH Commun. Math. Comput. Chem.*, 64(3):699–706, 2010. (Cited on page 31.)

[145] Kexiang Xu, Muhuo Liu, Kinkar Ch Das, Ivan Gutman, and Boris Furtula. A survey on graphs extremal with respect to distance-based topological indices. *MATCH Commun. Math. Comput. Chem.*, 71:461–508, 2014. (Cited on pages 31 and 122.)

[146] Guihai Yu and Lihua Feng. On connective eccentricity index of graphs. *MATCH Commun. Math. Comput. Chem.*, 69(3):611–628, 2013. (Cited on page 55.)

B

RAZŠIRJENI POVZETEK V SLOVENŠČINI

Socialna omrežja so že več desetletij predmet raziskav na različnih raziskovalnih področjih. Ob močni rasti internetnega omrežja v preteklih letih so se pojavila mnoga spletna socialna omrežja (na primer Facebook, Linkedin, Google+), poleg tega pa so postali dostopni podatki o drugih velikih omrežjih, kot je npr. omrežje soavtorstva člankov [109, 110, 118, 43].

Socialna omrežja ponavadi predstavimo z grafom, kjer so individualni ljudje predstavljeni kot vozlišča, razmerja med določenimi pari pa kot povezave v grafu. V večini omrežij so nekatera vozlišča ali povezave pomembnejša od drugih, zato je *centralnost* (ali *centralnost skupine*) pomemben koncept v raziskovanju teh socialnih omrežij [113, 52]. Glede na različne vrste omrežij ter njihovo uporabo je bilo definiranih več različnih centralnostnih indeksov. Med najbolj razširjene sodijo stopnja točke, vmesnostna centralnost, bližinska centralnost ter ekscentričnost. Poleg teh v disertaciji analiziramo še nekatere druge strukturne indekse, kot so Wienerjev indeks ter totalna razdalja. Vsi ti so definirani v razdelku 3.4 na strani 25.

Centralnostni indeksi skupin, vpeljani l. 1999 (Everett in Borgatti [46]), merijo pomembnost izbrane množice vozlišč v omrežju. Glede na vsako izmed razširjenih centralnostnih mer, lahko definiramo pripadajočo skupinsko centralnost. Več o tem si lahko preberemo v 7. poglavju na strani 89.

Naj omenimo še koncept Freemanove centralizacije [52], ki nam omogoča, da za izbrani strukturni indeks merimo relativno pomembnost znotraj omrežja, ki je primerljiva z rezultati iz drugih omrežij (več o tem v razdelku 3.3.2 na strani 24. Centralizacija strukturnega indeksa X je definirana kot

$$X_1(v, G) := \sum_{u \in V(G)} \big[X(v) - X(u)\big].$$

Disertacija je razdeljena v dva dela. V prvem delu se osredotočimo na ključne vsebine iz teorije grafov ter analize omrežij ter te zaporedoma predstavimo v 2. in 3. poglavju. Tu so med drugim definirani vsi zgoraj našteti strukturni indeksi, ter relevantne družine grafov, ki se pojavljajo v nadaljevanju. V drugem

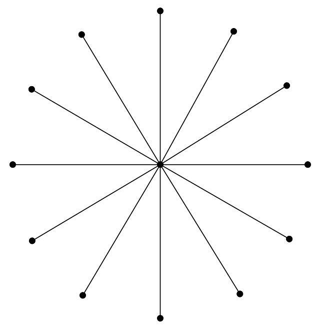


Figure B.1: Zvezda S_n na 13 vozliščih.

delu, tj. v poglavjih 4–10, predstavimo ključne rezultate v zvezi z zgoraj naštetimi strukturnimi indeksi.

Med področji, ki so nas zanimala, je tudi vprašanje iskanja grafov, ki maksimizirajo doseženo vrednost centralizacije nekega centralnostnega indeksa (t. i. ekstremalni grafi). Nekatere odprte hipoteze ter relevantne članke iz te teme lahko najdemo v Freeman [52], Borgatti and Everett [21], Everett and Borgatti [46], Sinclair [128], Everett et al. [48], Sinclair [129], Butts [28], Bell [17]. V [48] avtorji pokažejo, da je graf $H(v, n_0, n_1)$ (za definicijo glej razdelek 2.2 na strani 11) ekstremalni graf za centralizacijo vmesnostnega centralnostnega indeksa za družino povezanih dvodelnih grafov s fiksno velikostjo obeh biparticij. Hkrati avtorji domnevajo, da je isti graf ekstremalen tudi za centralizacijo bližine ter lastnih vektorjev. V poglavjih 4–6 razrešimo nekaj vprašanj o ekstremalnih grafih za centralizacijo bližine, ekscentričnost ter totalno razdaljo.

BLIŽINSKA CENTRALNOST DVODELNIH OMREŽIJ

V 4. poglavju se osredotočimo na ekstremalne grafe za centralnost bližine, ki meri, kako blizu je neko vozlišče v skupni razdalji do vseh ostalih vozlišč, natančneje

$$C_G(v) := \frac{1}{\sum_{u \in V(G)} d_G(v, u)}.$$

Centralizacija bližine je definirana kot

$$C_1(v, G) := \sum_{u \in V(G)} \big[C(v) - C(u)\big].$$

Graf *zvezda* S_{n-1}, kot je definiran v [134], je drevo na n točkah, kjer ima eno vozlišče stopnjo $n - 1$, vsa ostala vozlišča pa so listi (glej sliko B.1). Med vsemi povezanimi grafi na n vozliščih $\mathcal{G}_n$ je bližinska centralizacija maksimizirana prav v zvezdi.

Izrek 4.1 (stran 36)**.** *Naj bo G omrežje na n vozliščih. Potem velja*

$$C_1(u; S_{n-1}) \geqslant C_1(G),$$

kjer je u vozlišče z maksimalno stopnjo v grafu S_{n-1}.

V 4. poglavju potrdimo domnevo avtorjev Everett, Sinclair in Dankelmann [48] glede maksimiziranja bližinske centralizacije v dvodelnih omrežjih s podanimi velikostmi biparticij. Trdimo, da je največja vrednost centralizacije bližine (med vsemi dvodelnimi omrežji) dosežena, če lokalno maksimiziramo bližinsko centralnost v neki točki. Izkaže se, da je ekstremalna konfiguracija dosežena v korenskem drevesu globine 2, z dodatnim pogojem, da imajo vsi sosedje od korena skoraj enako stopnjo.

Izrek 4.3 (stran 40)**.** *Naj bo G dvodelno omrežje z biparticijama A_0 ter A_1, zaporedoma velikosti n_0 ter n_1. Potem, za $v \in A_0$, velja*

$$C_1(u; H(u; n_0, n_1)) \geqslant C_1(v; G).$$

EKSCENTRIČNOST OMREŽIJ S STRUKTURNIMI OMEJITVAMI

Ekscentričnost $e_G(v)$ vozlišča $v \in V(G)$ v povezanem omrežju G je maksimalna razdalja med u in v, kjer je u poljubno vozlišče v G, formalno

$$e_G(v) := \max\left\{d_G(v,u) \,:\, (u,v) \in V(G)^2\right\} \in \mathbb{N} \cup \{\infty\}.$$

V 5. poglavju med drugim določimo maksimizirajočo vrednost ekscentrične centralizacije ter najdemo nekaj maksimizirajočih omrežij za različne družine grafov. Označimo z $\mathcal{B}(n_0, n_1)$ družino vseh povezanih dvodelnih grafov z biparticijama velikosti n_0 ter n_1, z $\mathcal{T}_n$ pa družino vseh dreves na n točkah. V disertaciji najprej pokažemo, da se med ekstremalnimi grafi v družini $\mathcal{B}(n_0, n_1)$ vedno nahajajo tudi drevesa, ki jih v spodnjem izreku tudi karakteriziramo.

Izrek 5.3 (stran 58)**.** *Naj velja $n_0 \geqslant n_1 \geqslant 2$. Potem so grafi v $\mathcal{B}(n_0, n_1)^* \cap \mathcal{T}_{n_0+n_1}$ natanko drevesa iz $\mathcal{B}(n_0, n_1)$ z diametrom 4.*

Za družino vseh povezanih grafov na n vozliščih $\mathcal{G}_n$ ter povezanih grafov na m povezavah $\mathcal{G}_m$ naj omenimo spodnja nekoliko preprostejša rezultata.

Trditev 5.5 (stran 60)**.** Naj bo $n \geqslant 2$, $G \in \mathcal{G}_n$ ter $v \in V(G)$. Potem je $E_1(G, v) \leqslant \frac{n-1}{2}$, kjer velja enakost natanko tedaj, ko je v edino vozlišče stopnje $n-1$.

Od tod hitro sledi naslednja posledica, ki določi zvezdo za edini ekstremalni graf družine $\mathcal{G}_m$.

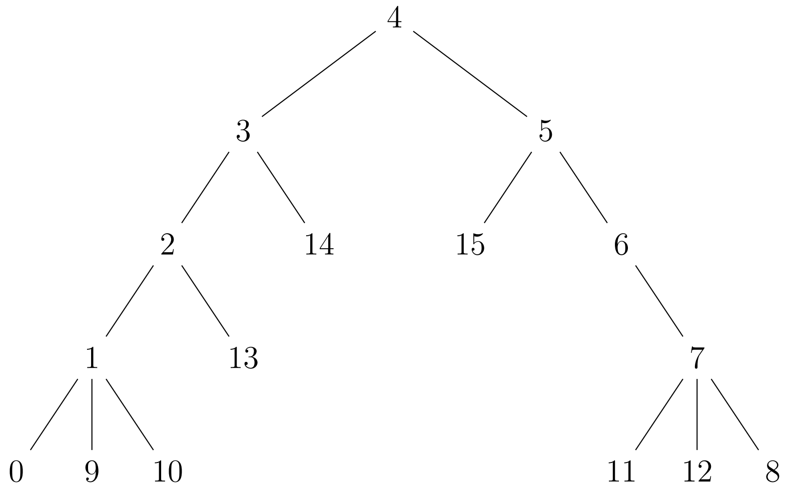


Figure B.2: Primer S-enumeracije drevesa s z maksimalno stopnjo 4, diametrom 8 ter 16 vozlišči.

Posledica 5.6 (stran 60)**.** *Če velja $m \geqslant 2$, tedaj je $E_1^*(\mathcal{G}_m) = \frac{m}{2}$ ter $\mathcal{G}_m^* = \{S_m\}$.*

Med drugim v podrazdelku 5.4.1 na strani 62 opišemo tudi nov način enumeracije vozlišč v poljubnem drevesu. Posebne lastnosti te enumeracije med drugim omogočijo enostavnejšo karakterizacijo nekaterih ekstremalnih grafov, ki jih omenimo kasneje.

Spodaj podajamo algoritmični postopek enumeracije vseh vozlišč nekega drevesa $T \in \mathcal{T}_n$ z diametrom d. Za koren drevesa vzemimo neko mediansko vozlišče. Enumeraciji omenjenega drevesa T pravimo S-enumeracija, če jo lahko generiramo s spodnjim postopkom:

- Za začetek vzemimo diametralno pot v T ter označimo pripadajoča vozlišča zaporedoma s števili $0, \dots, d$.

- Od sedaj naprej bomo označevali le vozlišča, katerih starši so že označeni. Izvedimo naslednjo zanko, pri kateri je vrednost spremenljivke i zaporedoma med 1 ter $\lfloor d/2 \rfloor$. V vsaki iteraciji ločeno izvedimo spodnji dve zanki:

  1. Vsakega neoznačenega otroka v, katerega starš je označen z i, ter njegovo poddrevo, oštevilčimo zaporedoma glede na algoritem iskanja v globino.
  2. Vsakega neoznačenega otroka v, katerega starš je označen z $d - i$, ter njegovo poddrevo, oštevilčimo zaporedoma glede na algoritem iskanja v globino.

Opazimo lahko, da takšna enumeracija ni enolično določena, saj je nedeterminističen že algoritem iskanja v globino, odvisna pa je tudi od izbire diametralne poti. Slika B.2 predstavlja primer S-enumeracije drevesa z maksimalno stopnjo 4, diametrom 8 ter 16 vozlišči.

Da lahko zapišemo izrek, ki karakterizira ekstremalne grafe za družino dreves fiksne velikosti n s podano maksimalno stopnjo Δ, potrebujemo nekaj dodatnih definicij. Za drevo T na n vozliščih s sodim diametrom $2k$, ki je korenjeno v svoji mediani m, pravimo, da je ekvivalentno z drevesom T', če lahko slednjega dobimo s "prerazporeditvijo" poddreves, korenjenih na neki fiksni razdalji od m, pri čemer ostaneta tako maksimalna stopnja kot diameter nespremenjena (natančno definicijo omenjene ekvivalenčne relacije najdemo na strani 62).

Za pozitivna števila Δ, k ter $n \geqslant \max\{\Delta + 1, 2k\}$ naj $F_{\Delta,k}$ predstavlja tako drevo z maksimalno stopnjo Δ ter diametrom $2k$, ki maksimizira število vozlišč, ter naj bo $F_{\Delta,k}(n)$ (enolično) poddrevo S-enumeracije drevesa $F_{\Delta,k}$, inducirano na vozliščih, označenih z $\{0, \ldots, n-1\}$. Naj $\mathcal{F}_{\Delta,k}(n)$ označuje družino vseh dreves, ki so ekvivalentna z $F_{\Delta,k}(n)$.

Ekstremalne grafe za družino dreves fiksne velikosti n s podano maksimalno stopnjo Δ karakteriziramo v naslednjem izreku.

Izrek 5.8 (stran 63)**.** *Naj bosta Δ ter n celi števili, tako da $3 \leqslant \Delta \leqslant n-3$. Potem je*

$$\mathscr{T}^*_{n,\Delta} = \mathcal{F}_{\Delta,k(n,\Delta)}(n),$$

kjer velja $k(n,\Delta) = \left\lceil \log_{\Delta-1}\left((n-1)\cdot\frac{\Delta-2}{\Delta}+1\right)\right\rceil.$

Ekstremalna drevesa seveda dosežejo največjo vrednost centralizacije ekscentričnosti v medianskem vozlišču, ki pa ni nujno stopnje Δ.

CENTRALIZACIJA TOTALNE RAZDALJE

Totalna razdalja vozlišča v je enaka vsoti vseh razdalj med v ter vsemi drugimi vozlišči v omrežju. Pri analizi centralizacije totalne razdalje v 6. poglavju določimo grafe na n točkah, ki dosežejo maksimalno ter minimalno vrednost le-tega indeksa. Izkaže se, da so maksimizirajoči grafi sestavljeni iz poti, ki je na enem koncu identificirana s kliko podobne velikosti. Te grafe natančneje opišemo v naslednji definiciji.

Definicija 6.5 (stran 78)**.** Za pozitivna cela števila a, b ter $c \leqslant b$, naj $\mathrm{PK}(a,b,c)$ predstavlja povezan graf na $a+b$ vozliščih, sestavljen iz poti P_a ter klike K_b, tako da je eno krajišče omenjene poti povezano s c vozlišči klike K_b. Graf $\mathrm{PK}(a,b,c)$ tako vsebuje $a - 1 + \binom{b}{2} + c$ povezav. Dva primera lahko najdemo na sliki B.3.

Figure B.3: Primera grafov PK(2, 2, 1) ter PK(2, 2, 2).

S pomočjo zgornje definicije formuliramo naslednji izrek.

Izrek 6.8 (stran 81)**.** *Grafi, ki v družini $\mathcal{G}_n$ maksimizirajo centralizacijo totalne razdalje, so izomorfni grafu*

- PK $\left(\frac{n}{2}, \frac{n}{2}, i\right)$*, če je n sodo število, pri čemer velja* $1 \leqslant i \leqslant \frac{n}{2}$*, ter*
- PK $\left(\lfloor\frac{n}{2}\rfloor, \lceil\frac{n}{2}\rceil, 1\right)$*, če je n liho število.*

V razdelku 6.3 pokažemo, da so grafi, ki minimizirajo centralizacijo totalne razdalje, sestavljeni iz treh poti podobne velikosti, ki imajo eno krajišče identificirano v skupni točki. Te grafe opišemo v naslednji definiciji. Naj $P(\alpha_1, \alpha_2, \alpha_3)$ predstavlja drevo na $1 + \alpha_1 + \alpha_2 + \alpha_3$ vozliščih, sestavljeno iz treh poti P_{α_1}, P_{α_2}, P_{α_3}, tako da je eno vsakega krajišča teh poti povezano z dodatnim vozliščem stopnje 3.

Izrek 6.12 (stran 86)**.** *Naj velja $n \geqslant 9$ ter naj bodo $\alpha_1 \leqslant \alpha_2 \leqslant \alpha_3$ pozitivna cela števila, tako da $\alpha_1 + \alpha_2 + \alpha_3 = n - 1$ ter $\alpha_3 - \alpha_1 \leqslant 1$. Potem so grafi, ki minimizirajo centralizacijo totalne razdalje v družini $\mathcal{G}_n$, izomorfni grafu $P(\alpha_1, \alpha_2, \alpha_3)$.*

SKUPINSKA CENTRALIZACIJA OMREŽNIH INDEKSOV

Centralnostni indeksi skupin, vpeljani l. 1999 (Everett in Borgatti [46]), merijo pomembnost izbrane množice vozlišč v omrežju. V 7. poglavju preučujemo skupinske indekse centralizacije ekscentričnosti, stopnje, ter vmesnostne centralnosti, glede na neko skupino vozlišč, velikosti k.

Za centralnostne indekse določimo nekatere ekstremalne grafe, ki dosežejo maksimalno vrednost centralizacije skupine. Za skupinsko centralizacijo stopnje v spodnjem izreku določimo, da je edini ekstremalni graf izomorfen zvezdi S_{n-1}.

Izrek 7.2 (stran 92)**.** *V družini $\mathcal{G}_n$ doseže zvezda S_{n-1} maksimalno vrednost skupinske centralizacije stopnje, ne glede na velikost skupine k.*

V nadaljevanju se osredotočimo na skupinsko centralizacijo parov točk, torej $k = 2$. Tu določimo največje dosežene vrednosti skupinske ekscentričnosti ter skupinske vmesnostne centralnosti, hkrati pa določimo tudi nekatere pripadajoče ekstremalne grafe.

Izrek 7.8 (stran 98)**.** *Naj bo $G \in \mathcal{G}_n$ graf, ki doseže maksimalno vrednost skupinske centralizacije ekscentričnosti, pri velikosti skupine $k = 2$. Potem pri pravilno izbrani dvojici vozlišč C velja:*

- *za $n = 3$ velja $G \simeq P_3$, ne glede na izbor C,*
- *za $n = 4$ velja $G \simeq S_3$, če je mediana $m \in C$,*
- *za $n = 5$ velja $G \simeq D(5, 3)$, če le C dominira G,*
- *za $n \geqslant 6$ velja $G = D(n, 5)$, če $C = \{p_0, p_3\}$.*

Naj bo $\mathrm{Tr}_n \in \mathcal{G}_n$ graf na n vozliščih, ki vsebuje natanko $n - 2$ različnih trikotnikov, ki pa si vsi delijo eno povezavo. Za skupinsko vmesnostno centralizacijo pri $k = 2$ dokažemo naslednje.

Izrek 7.12 (stran 103)**.** *V družini $\mathcal{G}_n$ je največja vrednost skupinske vmesnostne centralizacije, pri velikosti skupine $k = 2$, dosežena v grafu Tr_n.*

ALGORITMIČNI PRISTOP K SKUPINSKI CENTRALIZACIJI STOPNJE

Na problem določanja najboljše skupine v smislu skupinske centralizacije stopnje pri podanem omrežju G se osredotočimo tudi algoritmično. Pri podani velikosti skupine k problem zajema iskanje take množice vozlišč velikosti k, ki maksimizira skupinsko centralnost stopnje, ter nato računanje Freemanove centralizacije za to skupino, glede na skupinsko centralnost vseh ostalih skupin enake velikosti.

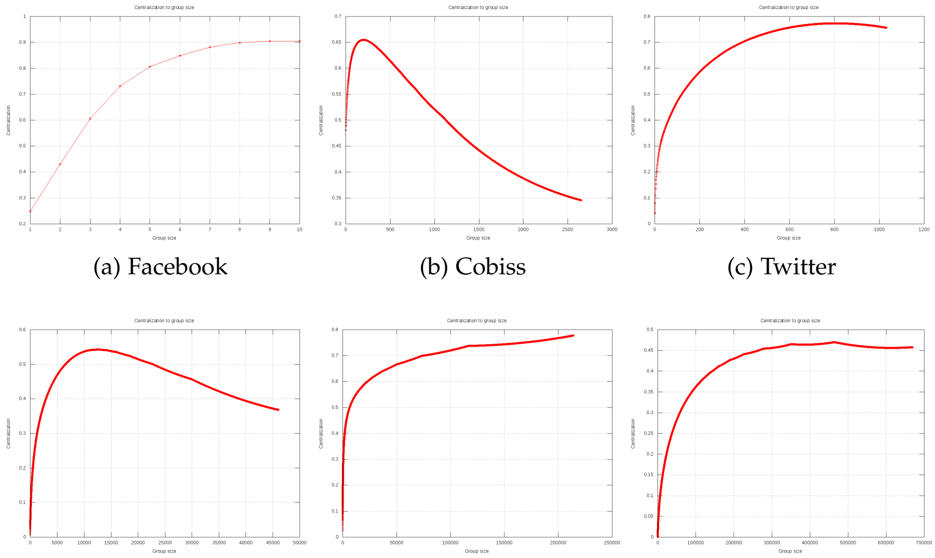
Prvi del omenjenega problema je $\mathcal{NP}$-težak, kar lahko opazimo, če ga prevedemo na problem maksimalne k-dominacije [107].

Trditev 8.4 (stran 110)**.** *Naj bo G vhodni graf na n vozliščih, ter naj bo $k < n$ pozitivno celo število. Problem, ki določi podmnožico $S \in \binom{V(G)}{k}$, ki maksimizira skupinsko centralnost stopnje je $\mathcal{NP}$-težak.*

Za drugi del problema (tj. računanje Freemanove centralizacije za podano množico vozlišč, glede na centralnost vseh ostalih k-teric) dokažemo, da se lahko izračuna v linearnem času.

Izrek 8.3 (stran 110)**.** *Naj bo G vhodni graf na n vozliščih, naj bo $k < n$ pozitivno celo število, ter naj bo S podana množica vozlišč velikosti k. Potem je skupinska centralizacija stopnje $\mathrm{GD}_1(S, G)$ enaka*

$$\mathrm{GD}_1(S,G) = \frac{\binom{n}{k} \cdot \mathrm{GD}(S,G) + \sum_{v \in V(G)} \binom{n-\deg(v)-1}{k} - n \cdot \binom{n-1}{k}}{(k+1) \cdot \binom{n-1}{k+1}},$$



(a) Facebook (b) Cobiss (c) Twitter

(d) Amazon (e) Youtube (f) Patents

Figure B.4: Grafična predstavitev Freemanove centralizacije skupinske centralnosti stopnje za šest omrežij različnih velikosti.

kar je izračunljivo v $O(n)$ korakih.

Glavni rezultat poglavja je polinomski algoritem (glej stran 112) z najboljšim možnim aproksimacijskim koeficientom, ki za vse smiselne velikosti k izračuna aproksimacijo centralizacijskih vrednosti v skupni časovni zahtevnosti $O(m + \gamma(G) \cdot \Delta(G)) \leqslant O(n^2)$. Omenjeni algoritem testiramo na šestih realnih omrežjih. Nekaj rezultatov je prikazanih v tabeli B.1. V rezultatih na sliki B.4 pri večini omrežij opazimo lastnost unimodalnosti (za parameter k), ki se lahko uporabi kot novi deskriptor za preučevanje velikih omrežij.

G	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$	$\mathrm{dc}(G)$	$\mathrm{GD}_1(S^*_{\mathrm{dc}(G)}, G)$	$\mathrm{GD}(S^*_{\mathrm{dc}(G)}, G)$
facebook	4,039	88,234	10	0.905393	4,029
cobiss	25,301	316,587	204	0.654997	19,635
twitter	81,306	1,768,149	803	0.773615	78,811
amazon	403,394	3,387,388	12,810	0.542647	320,133
youtube	1,134,890	2,987,624	214003	0.777639	920,887
patents	3,923,922	16,522,438	464,298	0.470009	3,105,485

Table B.1: Nekaj statističnih podatkov iz analiziranih realnih mrež.

O INTERVALU WIENERJEVEGA INDEKSA

Wienerjev indeks $W(G)$ grafa G je enak vsoti razdalj med vsemi pari vozlišč v G. Z $W[\mathcal{G}_n]$ označimo množico vseh vrednosti Wienerjevega indeksa za družino povezanih omrežij na n vozliščih, pri čemer največji neprekinjen interval iz $W[\mathcal{G}_n]$ označimo z W_n^{int}. Iz spodnje trditve sledi, da je trivialna zgornja meja velikosti obeh intervalov enaka kvečjemu $\binom{n+1}{3} - \binom{n}{2} + 1$.

Trditev 9.4 (stran 122)**.** *Za družino $\mathcal{G}_n$ velja*

$$\min\left(W\left[\mathcal{G}_n\right]\right) = \binom{n}{2} \ \textit{ter} \ \max\left(W\left[\mathcal{G}_n\right]\right) = \binom{n+1}{3}.$$

Spodnja in zgornja meja sta zaporedoma doseženi v K_n ter P_n.

V 9. poglavju pokažemo, da je W_n^{int} smiselno definiran ter se, za $n \geqslant 25$, začne v vrednosti $\binom{n}{2}$. Poleg tega na konstruktiven način postavimo spodnjo mejo za velikost W_n^{int} ter $W[\mathcal{G}_n]$, ki je asimptotično blizu omenjeni zgornji meji, tj. najdemo grafe, ki s svojimi Wienerjevimi vrednostmi nepretrgoma večinoma pokrijejo interval med $\binom{n}{2}$ ter $\binom{n+1}{3}$. Natančneje pokažemo naslednje.

Izrek 9.11 (stran 127)**.** *Naj bo $W_n^{\text{int}} = [a, b]$ največji interval zaporednih celih števil, ki je v celoti vsebovan v $W[\mathcal{G}_n]$. Potem velja*

$$a = \binom{n}{2} \quad \textit{and} \quad b \geqslant \frac{1}{6}n^3 - \frac{5}{2}n^2 + O\left(n^{3/2}\right).$$

Med drugim sledi $\left|W_n^{\text{int}}\right| = \frac{1}{6}n^3 + O\left(n^2\right)$.

NADALJNJE DELO

Zgornje rezultate predstavimo v ločenih poglavjih, ter jih zaključimo z morebitnimi idejami za prihodnje delo ter odprtimi domnevami. V zaključnem poglavju vključimo kratek povzetek tekočega dela v zvezi z vmesnostno centralizacijo, izpostavimo nekatere predstavljene rezultate ter pregledno povzamemo nekatere odprte domneve na področju ekstremalnih grafov.

C

IZJAVA O AVTORSTVU

Podpisani Matjaž Krnc izjavljam:

- da sem doktorsko disertacijo z naslovom *Centrality measures of large networks* izdelal/a samostojno pod mentorstvom prof. dr. Riste Škrekovski,
- da se vsi soavtorji skupnih člankov navedenih na strani xi strinjajo z objavo rezultatov v tej disertaciji in
- da Fakulteti za matematiko in fiziko Univerze v Ljubljani dovoljujem objavo elektronske oblike svojega dela na spletnih straneh.

Ljubljana, 18. 9. 2015

Matjaž Krnc