Evaluating Group Degree Centrality and Centralization in Networks

Mario Karlovčec[1], Matjaž Krnc[2] and Riste Škrekovski[3]
E-mail: mario.karlovcec@ijs.si, matjaz.krnc@gmail.com, skrekovski@gmail.com

[1]Artificial Intelligence Laboratory, Jozef Stefan Institute and
Jozef Stefan International Postgraduate School Jamova 39, 1000 Ljubljana, Slovenia

[2]University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Koper, Slovenia

[3]Faculty of information studies, Novo Mesto, Slovenia

Keywords: degree centralization, group centrality, network analysis

Received: November 8, 2021

Given a network G, the importance of groups can be modeled by group centrality measures. Freeman's centralization is a way to normalize any given centrality or group centrality measure, which enables us to compare individuals or groups from different networks. We focus on a degree-based measure of group centrality and centralization, presented by Krnc and Škrekovski (2020). We describe its efficient implementation and study the behaviour of various real-world networks within this context. We conclude that very small groups, as well as very big ones, are not very central, i.e. as the group is growing, its value is increasing but, at some point it starts decreasing. Such unimodular behaviour is confirmed by our analysis of of group degree centralization of six real-world networks. At the end, we provide some challenges for future work.

Povzetek: Glede na podano omrežje G lahko pomembnost skupin modeliramo s pomočjo mer skupinske centralnosti. Freemanova centralizacija predstavlja način normalizacije za poljubno obstoječo mero centralnosti, kot tudi za poljubno obstoječo mero skupinske centralnosti. Omenjena normalizacija omogoča primerjavo posameznikov oz. skupin različnih omrežij. Članek se osredotoča se na mero centralnosti in centralizacije skupine, ki temelji na stopnji, kot je predstavljeno v Krnc in Škrekovski (2020). Opišemo njegovo učinkovito izvajanje in preučimo obnašanje različnih omrežij v resničnem svetu, glede na to mero. Ugotavljamo, da zelo majhne skupine, kot tudi tudi zelo velike, nimajo velike centralnosti — ko skupina raste, njena vrednost narašča, a na neki točki začne upadati. Takšno unimodularno *vedenje potrjuje naša analiza centralizacije skupinske stopnje šestih omrežij iz realnega sveta. Na koncu podamo nekaj izzivov za prihodnje delo.*

1 Introduction

In most networks some vertices are more central than the others. To model this intuitive feeling, centrality indices were introduced. The first mathematical concept of centrality of graphs was introduced almost 150 years ago by Jordan [13]. 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 areas of science.

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. In the paper, we will therefore freely interchange the terms vertex/node and graph/network, without any meaningful difference. Various vertex-based measures of centrality have been proposed to determine the relative importance of a vertex within a graph.

Arguably, the most common branch of centrality indices is based on the distance between the nodes of the network. Some of the standard centrality indices from this branch are degree, betweenness, closeness and eccentricity. Among other measures of node centrality, a few better known in network analysis are: eigenvector centrality, Google PageRank, Katz centrality, Alpha centrality, and others. For detailed definitions and discussions on various centrality indices, we refer the reader to [5, 1, 2, 14, 22, 23].

In large networks, measuring a vertex centrality with respect to the whole network is often not relevant, nor computationally feasible. In this sense, another concept of vertex centrality with respect to some subset of vertices has been introduced and studied throughout the last decade. The *personalization*, introduced in 2003 (see [25]), is a measure that shows how central an individual is according to a given subset R (group of important people) in a given social network. In 2005, the *subgraph centrality* [6] was introduced, which characterizes the participation of each node in all subgraphs in a network and is calculated from the spectra of the adjacency matrix of the network. In the same year, Everett et al. [8] introduced the *core centrality* measure, where they evaluate the extent to which a network revolves around a core group of nodes. Finally, very recently Bell [3] introduced the concept called the *subgroup centrality*, where centrality (of one vertex) is calculated only on a restricted set of vertices. Arguably, the most basic measure is the degree of a vertex, which we study in this paper.

In 1999, Everett and Borgatti [7] 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, men or women?" With these measures we can also solve the inverse problem: given a network of ties among organization members, how can we form a team that is maximally central? In [7], the authors introduced group centrality for measures of degree, closeness and betweenness centrality, which we use in this paper. In 2006, Borgatti introduced an important group centrality measure (usually called KPP) that is motivated by the *key players problem* (see [4]). In his paper he focused on finding a set of vertices for the purpose of optimally diffusing information through the network by using selected vertices as seeds, or to maximally fragment 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 such as graph fragmentation and information entropy to define KPP centrality.

In his study, Freeman [10] realized that despite all of the vertex-centrality indices defined up to that point, there was a need for a normalization which could measure a *relative* importance of a given vertex in a network and would be based on any chosen centrality index. Hence, he defined a *centralization* measure based on normalized variance in vertex centrality of any chosen centrality measure, with an aim to allow a comparison of whole networks on the basis of their highest vertex-centralization scores. One may also consider his approach as another type of vertex-centrality which measures the extent of how some vertex in a network stands out from others in terms of a given centrality index. This is useful since it arguably allows us to compare the centralization scores of nodes that belong to different networks.

In the same paper Freeman remarked that the centralizations of degree centrality, betweenness centrality and closeness centrality achieve their maximum if and only if G is a star. The statement was later proved in detail by Everett, Sinclair and Dankelmann [9]. In order to compare centralization values of graphs with different sizes, in the definition of centralization, Freeman used a normalized formula, where the normalizing divisor is based on the theoretically largest centrality variance in any graph from a given class of graphs [10].

1.1 Related work

Following Freeman's approach, the group centralization notion was introduced in [16], where the authors studied some extremal graphs regarding several group centrality measures, including group degree centralization. They showed that the maxumum value of group degree centralization is attained in a star graph, with centrality value of $(k+1)\binom{n-1}{k+1}$, see Proposition 2.

Our work is related with the domination theory. Among the relevant papers Miyano et al. [21] discussed the problem of finding the best group for the so called *k-vertex maximum domination problem* (or k-maxVD, in short). Although they claim that k-maxVD is a new variant of vertex-domination problem, it is in fact equivalent to maximizing the group degree centrality (introduced 12 years earlier by Everett and Borgatti [7]), with the score further increased by a constant k. In the paper authors reduce the problem to a *Maximum Coverage problem*, which is nicely discussed and analyzed by Vorha and Hall [24], or by Hochbaum and Pathria [12].

In [15] authors provide a conceptual description of a greedy procedure for estimating maximal k-group degree centralization values, for any value of k. Those results are based on a greedy appriximation algorithm for MAX COVERAGE or MAX k-VERTEX DOMINATION problems (see [21, 12]). While this greedy approach gives approximation ratio of $1-1/e$ for both mentioned algorithms, it is shown in [15] that any such constant is not attainable for the k-group degree centrality.

The historical overview of the above-mentioned results is illustrated in Table 1, while the approximability and time complexity of Algorithm 1, is described as follows.

Theorem 1 (Krnc and Škrekovski [15]). *Given a graph G, the greedy algorithm for k-group degree centrality over all set sizes $1 \leq k \leq n$ altogether runs in linear time and achieves the k-group degree centrality value of at least $(1-1/e)(w^*-k)$, where w^* is the maximizing k-group degree centrality of G.*

The rest of this paper is structured as follows. In Preliminaries we provide notations and definitions from [16, 15] that we use. We give definitions for Freeman centralization of group degree centrality and centralization. And revisit some relevant results from the field. In the next section we develop an efficient greedy algorithm for finding a group with approximate maximal degree centralization over all group sizes k from 1 to n and altogether runs in linear time. We describe the procedure in detail and provide complexity analysis of the algorithm. In Section 4 we continue our exposition by describing the behaviour of the implemented algorithm when used on some real-world complex networks. We describe the experiments made with six networks, ranging up to 3.9 million vertices and 16 million edges. We present the datasets used and discuss the results. In experiments we observe that, in the examples studied, there seems to be only one candidate for optimal value of k in each network. In other words, the shape of a function

Combinatorial Optimization

1993	Vorha and Hall [24]	Maximal covering location problems and analysis
1998	Hochbaum and Pathria [12]	
2011	Miyano and Ono [21]	Max k-vertex domination

Network Analysis

1869	Jordan [13]	Network centrality
1979	Freeman [10]	Network centralization
1999	Everett and Borgatti [7]	Group degree centrality
2004	Everett et al. [9]	Extremal networks for degree centrality
2015	Krnc and Škrekovski [16]	Group degree centralization

2020	Krnc and Škrekovski [15]	Conceptual description of algorithm
Present paper		Big data empirical analysis

Table 1: The milestones from the fields of Network Analysis and Combinatorial Optimization, leading to the present paper.

of group degree centralization over a group size seems to contain only one evident local minimum. We conclude the paper with remarks and future work.

2 Preliminaries

Let $\mathcal{G}_n$ be a family of all graphs on n vertices, let $G \in \mathcal{G}_n$ be a graph on m edges, and let $S \subseteq V(G)$. According to [7], the group degree centrality is defined as

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

Given a graph G and integer 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. $\mathrm{GD}_G(S_k^*) = \max_{S\in\binom{V(G)}{k}} \mathrm{GD}_G(S)$. Whenever the graph G is known from the context, we omit the subscript from the notations of centrality or centralization.

According to [10, 16], $\mathrm{GD}_1(S, G)$ stands for the group degree centralization. Define $k := |S|$, and observe that $\mathrm{GD}_1(S, G)$ is equal to

$$\frac{\sum_{S'\in\binom{V(G)}{k}} (\mathrm{GD}_G(S) - \mathrm{GD}_G(S'))}{\max_{H\in\mathcal{G}_n} \sum_{S''\in\binom{V(H)}{k}} (\mathrm{GD}_H(S_k^*) - \mathrm{GD}(S''))}. \qquad (1)$$

According to Freeman [10], the denominator is needed to efficiently normalize centralization to the interval $[0, 1]$, for better relative comparison. Clearly $\mathrm{GD}_1(S, G)$ is maximized whenever $\mathrm{GD}(S, G)$ is maximized, and by Krnc et al. [16] we have that the maximum value of the denominator corresponds to the star graph S_n and a maximizing set S_k^* corresponds to any k-set containing the center of the star. In Fig. 1 the calculation of the group degree centralization is demonstrated.

Denote the *maximizing group size* $\mathrm{dc}(G)$ to be the positive integer for which $S^*_{\mathrm{dc}(G)}$ achieves the maximum value of the 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 the minimum cardinality of any set that dominates a graph G (also known as the *domination number*). The notation $\Delta(G)$ stands for the highest vertex-degree, i.e. $\Delta(G) = \max_{v\in V(G)} \deg_G(v)$. A function f is said to be *unimodal* if it contains only a single local maximum in f.

2.1 Evaluating degree centralization

The goal of this section is to optimize the procedure of calculating the group degree centralization for a given graph and an input integer k. To this end we need to revisit certain important parts of its definition which will be later used in our procedures.

In [16] authors describe the denominator of (1) in a closed form, as follows.

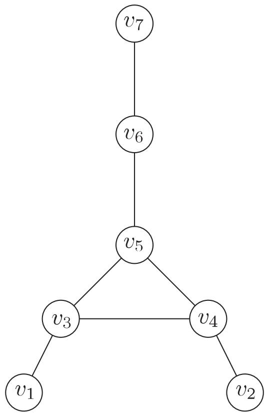
Proposition 2 (Krnc and Škrekovski [16])**.** *Let G be a star on n vertices with the center c, and let $S^* \in \binom{V(G)}{k}$ such that $c \in S^*$. Then*

$$\sum_{S'\in\binom{V(G)}{k}} [\mathrm{GD}(S^*) - \mathrm{GD}(S')] = (k+1)\binom{n-1}{k+1}.$$

In order to compute the sum $\sum_{S'\in\binom{V(G)}{k}} \mathrm{GD}(S', G)$ from (1) efficiently, we need the following claim.

Proposition 3 (Krnc and Škrekovski [15])**.** *Let G be a graph on n vertices, and let $k \le n$ be a positive integer. It holds that $\sum_{S'\in\binom{V(G)}{k}} \mathrm{GD}(S', G)$ is equal to*

$$n \cdot \binom{n-1}{k} - \sum_{v\in V(G)} \binom{n - \deg(v) - 1}{k}.$$



S	$GD(S)$	$GD_1(S)$
$\{v_1\},\{v_2\},\{v_7\}$	1	−0.23
$\{v_6\}$	2	0
$\{v_3\},\{v_4\},\{v_5\}$	3	0.23
$\{v_6,v_7\}$	1	−0.65
$\{v_1,v_2\},\{v_1,v_3\},\{v_1,v_7\},$ $\{v_2,v_4\},\{v_2,v_7\}$	2	−0.30
$\{v_1,v_4\},\{v_1,v_5\},\{v_1,v_6\},\{v_2,v_3\},$ $\{v_2,v_5\},\{v_2,v_6\},\{v_3,v_4\},\{v_3,v_5\},$ $\{v_4,v_5\},\{v_5,v_6\},\{v_5,v_7\}$	3	0.05
$\{v_3,v_6\},\{v_3,v_7\},\{v_4,v_6\},\{v_4,v_7\}$	4	0.40

Figure 1: The Eiffel-tower graph and the calculation of its group degree centralization for groups of sizes one and two.

In particular, the sum $\sum_{S'\in\binom{V(G)}{k}} \mathrm{GD}(S',G)$ *can be computed in* $O(n)$.

The results from Propositions 2 and 3 can be joined to further develop (1), i.e. $\mathrm{GD}_1(S,G)$ evaluattes to

$$\frac{\binom{n}{k}\cdot \mathrm{GD}(S,G)+\sum_{v\in V(G)}\binom{n-\deg(v)-1}{k}-n\cdot\binom{n-1}{k}}{(k+1)\cdot\binom{n-1}{k+1}}, \tag{2}$$

which can be computed in $O(n)$ steps.

It is easy to see that finding the best possible group S^* that maximizes the group degree centrality (and hence the group degree centralization) can be computed in $O(n^{k+1})$, traversing over all k-tuples and computing group degree centrality at each step. Note that, since an input integer k is in the exponent, this straightforward approach is far from polynomial.

In fact it is easy to see that the mentioned problem is $\mathcal{NP}$-hard. This can be done reducing a well-known $\mathcal{NP}$-problem of determining the existence of a k-dominating set to our problem of finding S^*_k. Let us assume that there exists a 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 a 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 polynomial time $\mathcal{O}(nk)$, it is clear that the set S^*_k provides us an answer regarding the existence of a k-dominating set. This trivially implies:

Proposition 4. *The problem that determines a set S^*_k for a given input graph G and an integer k is $\mathcal{NP}$-hard.*

In the last section we present an efficient algorithm that achieves the best possible linear-time approximation for calculating group degree centrality scores for all group sizes.

3 Algorithmic approach

For reasons described in Introduction, finding the group with the biggest degree centralization may be needed for many real-world networks, which can be very large in some cases (the largest network from the experiments contains 16 million edges). As shown in (2), calculating the group degree centralization for a given set S can be computed efficiently, while on the other hand, finding a maximizing set is $\mathcal{NP}$-hard (see Proposition 4). In this section we present an efficient implementation and demonstration of calculating an *estimate* of the group degree centralization for a given network. In particular we implement and study in detail a greedy approximate algorithm presented in [15], for finding a group with maximal degree centralization.

To calculate the group degree centralization efficiently, in [15] authors use a greedy algorithm for Maximum Coverage Problem which is a polynomial time $(1-1/e)$-approximation algorithm, see [12] or [21]. Here we describe the implementation of such algorithm in more detail, while retaining the same time efficiency. Particularly, more emphasis is given on efficient calculation of group centralization, as not all details are described in [15]. An implementation of the procedure that calculates an approximation for the group with the biggest degree centralization, for all meaningful group sizes, is given by Algorithm 1. Let us now describe the parts of the algorithm.

3.1 Algorithm description

We start with $k=0$ and in each step of the main while loop increment the size of group S. Every time when k increases, we add some greedily chosen vertex to the set S and remove it from G while maintaining some dictionaries that we use ($contribution$ and $histogram$, in particular). Note that by the data structure of a *dictionary* D we mean that D is a set of keys with additionally defined values $D\left[i\right]$ for each $i\in D$.

In the first phase of Algorithm 1 (throughout lines 1–12), we construct graph G from an input graph G' by orienting

Algorithm 1 Finding a group with the maximal degree centralization

Input: a graph G'.

Output: a list $centralization$ of group degree centralization scores, where $centralization[i]$ is an approximation of $\max_{S\in\binom{V(G')}{i}} \mathrm{GD}(S, G')$.

```
$n \leftarrow |V(G')|, S \leftarrow \emptyset$                              ▷ Centrality variables initialization.
$dominated \leftarrow \emptyset, histogram \leftarrow \emptyset$
$G \leftarrow$a directed instance of graph $G'$
**for all** $v$ in $V(G')$ **do**
    add $v$ to $histogram[\deg_{G'}(v)]$
    $dominated[v] \leftarrow$ False
$centralization \leftarrow \emptyset$                                   ▷ Centralization variables initialization.
$A \leftarrow 1/(n-1)$
$C \leftarrow n/(n-1)$
**for all** $i \in histogram$ **do**
    $sum[i] \leftarrow 1/(n-1)$
    $degDistribution[i] \leftarrow |histogram[i]|$
**for all** $0 \leq k \leq n$ **do**                                      ▷ Main loop
    $v \leftarrow$ any vertex from $histogram$ with the highest contribution
    $S \leftarrow S \cup v$
    $k \leftarrow k + 1$
    $\mathcal{GD} \leftarrow \mathcal{GD}+$CONTRIBUTION($v$)
    **for all** $u \in N_G^-(v)$ **do**
        DECREASECONTRIBUTION($u$)
    **for all** $u \in N_G^+(v)$ **do**
        DECREASECONTRIBUTION($u$)
        $dominated[u] \leftarrow$ True
        **for all** $w \in N_G^-(u)$ **do**
            DECREASECONTRIBUTION($w$)
            $E(G) \leftarrow E(G) - uw$
    $G \leftarrow G - v$
    UPDATECENTRALIZATIONVARIABLES()
    $centralization[k] \leftarrow A \cdot \mathcal{GD} + B - C$           ▷ Computing centralization.
**return** $centralization$
```

every edge of G' in both directions. We also initialize the starting values of all dictionaries and other variables. In the *main loop* (lines 13–28) we first chose the vertex v to add to our group S, update variables S and k accordingly (lines 14–16), and then calculate the group degree centrality for the increased set S (in line 17). Then, throughout the lines 18–26 we update the dictionary $histogram$, and also remove v and some additional directed edges from the graph. Finally, in lines 27 and 28, we update some centralization variables and calculate the Freeman centralization of the centrality from the line 17. We now present some details of how we maintain some values and graph properties.

For each group, we efficiently calculate the group degree centralization using Propositions 2 and 3. 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 the algorithm. Let S be the group of vertices whose group degree centralization we are calculating, and set $k := |S|$. Note that we initially have $\deg^-_G(v) = \deg^+_G(v) = \deg_{G'}(v)$ for all vertices in the network. Before the beginning of each iteration of the *main loop*, the existence of a directed edge uv means that

- in the initial graph G', we have $uv \in E(G')$,
- neither of v, u is a member of S, and
- in the initial graph G', vertex v is not connected with any vertex from S, i.e. $v \notin \cup_{v\in S} N(v)$.

For any vertex $v \in V(G) \setminus S$, we define the *contribution* of v to be the value $\mathrm{GD}\left(S \cup \{v\}, G\right) - \mathrm{GD}\left(S, G\right)$ and observe that

$$\mathrm{GD}\left(S \cup \{v\}, G\right) - \mathrm{GD}\left(S, G\right)$$

evaluates to

$$\begin{cases} \deg^+_G(v), & \text{if } v \text{ is not dominated,} \\ \deg^+_G(v) - 1, & \text{otherwise.} \end{cases}$$

The calculation of the value of the contribution is done by a short function $contribution$ (see Algorithm 3). As different nodes have various contributions, we define the dictionary $histogram$, initialized in line 5 of Algorithm 1, where the keys are all possible values of contribution (for any key i, it clearly holds that $-1 \le i \le \Delta(G)$), and the values are unordered sets of nodes with corresponding contributions. While the dictionary $histogram$ is initially indeed a degree histogram, we update it with each modification of variables G or S. The goal of the algorithm is to calculate the value of (2) for each S. We implement this by introducing variables $A, B, C, \mathcal{GD}$, each of them assigned to a different part of the expression (2), in particular,

$$\begin{aligned} A &= \frac{n}{(n-k)\,(n-k-1)}, \\ \mathcal{GD} &= \mathrm{GD}(S,G), \\ B &= \frac{(n-k-2)!}{(n-1)!} \sum_{v\in V(G)} \frac{(n-\deg(v)-1)!}{(n-k-\deg(v)-1)!}, \\ C &= \frac{n}{n-k-1}. \end{aligned}$$

Hence, for algorithmic purposes (see Propositions 2 and 3, and Eq. (2)), we may write

$$\mathrm{GD}_1(S,G) = A \cdot \mathcal{GD} + B - C.$$

Clearly, whenever the group S increases, the values of $A, B, C, \mathcal{GD}$ also change. To handle the change of variable B, we introduce dictionaries $degDistribution$ and sum. The keys of both are all possible degrees of the vertices in G, and the values are defined as

$$\begin{aligned} sum[i] &= \frac{(n-i-1)!}{(n-k-i-1)!} \cdot \frac{(n-k-2)!}{(n-1)!}, \\ degDistribution[i] &= |\{v : deg_G(v) = i\}|. \end{aligned}$$

Algorithm 2 Updating variables A, B, C and sum.

```
1: function UPDATECENTRALIZATIONVARIABLES
2:     $A \leftarrow \frac{n}{(n-k)(n-k-1)}$
3:     $C \leftarrow \frac{n}{n-k-1}$
4:     for all $i$ in $sum$ do
5:         $sum[i] \leftarrow sum[i] \cdot \frac{n-i-k-1}{n-k-2}$
6:     $B \leftarrow \sum_i sum[i] \cdot degDistribution[i]$
```

Note that sum needs 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 using the 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 2.

Note that while the graph is changing and the group S is increasing, the contributions of the remaining vertices also change. While the initial contribution of a vertex v is equal to its degree $\deg_G(v)$, during the main loop the contribution of v may decrease by one several times. We handle these changes by defining a function $decreaseContribution(v)$, see Algorithm 3.

4 Experiments

In this section we describe the experiments made with six real-world networks. We present the datasets used and describe their domain. We implement the algorithm in C++ and Python with the help of the libraries “SNAP” and “SNAP.Py”, respectively. Finally, we discuss the results and establish the correlation of the results with the unimodality law.

Algorithm 3 Functions $contribution$ and $decreaseContribution$. The former outputs the contribution of v while the latter refreshes the dictionary $histogram$ whenever the contribution of v decreases by one.

```
function CONTRIBUTION($v$)
    if $v$ is dominated then
        return $\deg^{+}_{G}(v) - 1$
    else
        return $\deg^{+}_{G}(v)$
function DECREASECONTRIBUTION($v$)
    $c \leftarrow$ CONTRIBUTION(v)
    $histogram[c] \leftarrow histogram[c] \setminus \{v\}$
    $histogram[c-1] \leftarrow histogram[c-1] \cup \{v\}$
```

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 to more than 16 million edges.

Facebook is the smallest dataset in experiments containing 4039 nodes and 88,234 edges. It is anonymized data collected by survey participants using a Facebook application with ten ego networks combined. The dataset was generated to study social circles in ego networks [19]. Since it is generated by combining 10 ego networks, we expect its maximal centralization results to be relatively high. 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 by using the database maintained by ARRS (Slovenian Research Agency) and IZUM (Institute of Information Science, Maribor, Slovenia). Twitter dataset contains a graph of followers, with 81,306 nodes and 1,768,149 edges. The dataset was collected from public sources for the purpose of analyzing the social circles [19]. Amazon dataset is a graph of frequently co-purchased products based on the Amazon website in June, 2003. The mentioned graph has 403,394 nodes and 3,387,388 edges and was generated for the study of viral marketing dynamics [17]. YouTube dataset contains a graph of user subscriptions. The graph contains 1,134,890 nodes and 2,987,624 edges. The dataset was prepared by Mislove et al. [20]. Patents dataset is a citation graph of patents granted between 1975 and 1999. The graph contains 3,923,922 nodes (patents) and 16,522,438 edges (citations). The dataset was generated for the purpose of studying graph evolution [18], using the U.S. patent dataset maintained by the National Bureau of Economic Research [11].

4.2 Experimental results

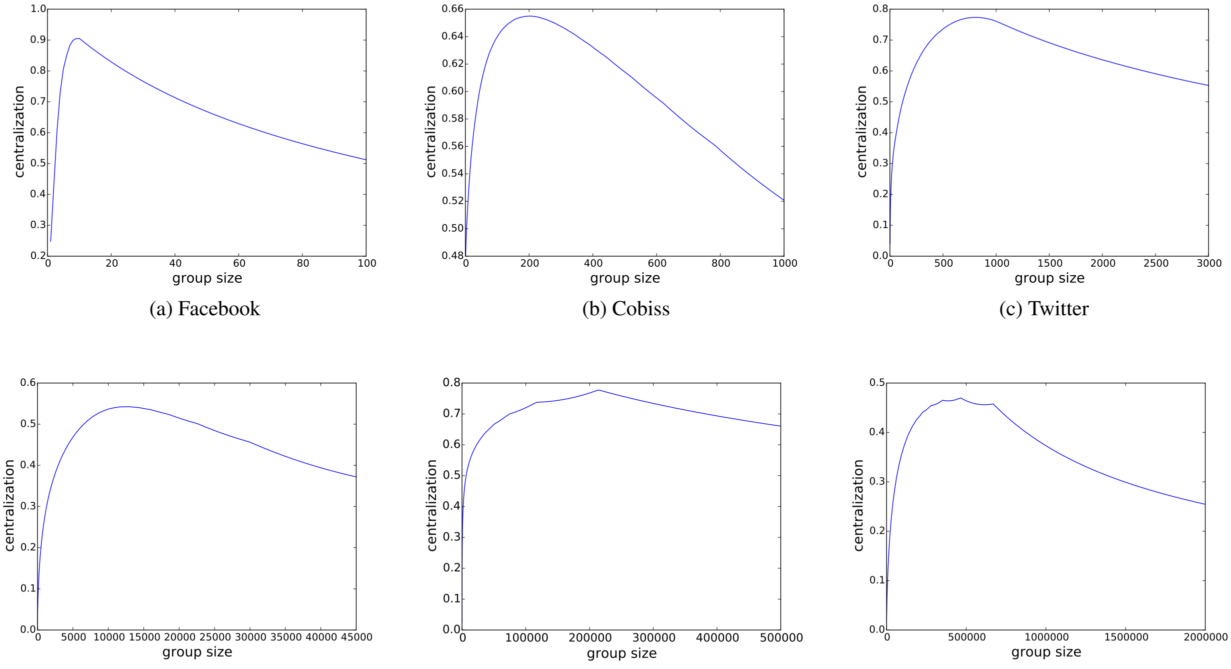
Figure 2 shows the values of the centralization with different sizes of the group, while Table 2 gives precise values of results and optimal dataset sizes. In Cobiss network (Figure 2b) the maximal centralization is attained by a relatively small group size, and after that point adding members to the group causes a drastic decrease of the centralization. Amazon (Figure 2d) has a similar shape, but increasing and decreasing of centralization is less intensive. In YouTube network (Figure 2e), a relatively small group size has a high degree centralization. Further increase of the group size slowly increases centralization, which is maximal only after the group dominates all the nodes. In Facebook network (Figure 2a) the maximal centralization is also achieved when the group dominates all the nodes. Patents network (Figure 2f) is similar to YouTube, but the maximal centralization is achieved before the nodes are dominated by the group. In Twitter (Figure 2c), network centralization increases up to the maximal point and then decreases with the same intensity.

By its construction, the group degree centralization is comparable among various groups from different networks. Among the analyzed networks, one can observe that the biggest centralization score (0.9) is attained by the Facebook network, with the corresponding group of size 10. This result was expected, since 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 by 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 the Patents network, which is also the largest and the sparsest network among our experiments.

By looking at the centralization values for all networks in the experiments and thousands of different sizes of groups, we observed a correlation between the plots in Figure 2 and the unimodality law. Indeed, the shape of most of the plots in Figure 2 is unimodal, i.e. it is monotonically increasing up to a maximizing group size, while at the value $\mathrm{dc}(G)$ the plot begins to have a negative slope. As discussed above, this confirms the natural intuition regarding the group centrality scores, mentioned in unimodality law. While this property is certainly not present in all networks, as it is easy to find many graphs that do not follow the unimodality law, we believe that many of the real-world networks should indeed have this unimodality property and that it should be studied further.

5 Discussion

With respect to the important branches of the centrality theory mentioned above, we seek to identify the most 'central' group of nodes in a network, without assuming that we know in advance its cardinality. For a given network G,



(a) Facebook (b) Cobiss (c) Twitter

(d) Amazon (e) YouTube (f) Patents

Figure 2: The graphic representation of the Freeman centralization of the group degree centrality of six networks with different sizes of groups.

G	$\|V(G)\|$	$\|E(G)\|$	$\mathrm{dc}(G)$	$\mathrm{GD}_1(S^*_{\mathrm{dc}(G)}, G)$	$\mathrm{GD}(S^*_{\mathrm{dc}(G)}, G)$
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 2: Some statistics and centrality results from our experiments.

we mainly deal with three correlated questions:

Q1. For a fixed k, which k-subset S of members of G represents the most influential group?

Q2. Among all possible values of k find the one for which the corresponding set S from Q1 is most influential.

Q3. How to efficiently compute both k and S from Q2 in real-world scenarios?

Group centrality measures often have the property that, for a given group S, one can find a person $x \notin S$, such that $S \cup x$ has higher centrality score than S. While this intuition applies to most of known group centrality indices, it does not provide the desired result, as the best group will in most cases correspond to the whole network. In the sense of Q2, it seems natural that groups of very small sizes should not be very influential, and we would expect the influence to grow while the size of the group increase. Stating this more explicitly, we have:

Law of small groups. *For most small groups S and any person $x \notin S$, the extended group $S \cup x$ should be more central than S.*

However, at some point we would like to notice that the further increase of the group size starts decreasing its score in the sense of given group centrality measure, as the group is too big and consequently less tractable. The intuitive phenomena mentioned above is expressed by the following rules:

Law of large groups. *For most large groups S and any person $x \notin S$, the extended group $S \cup x$ should be less central than S.*

Of course above two laws are not meant to hold strictly, and are therefore stated in a rather mathematically non-precise way. Anyway, in order to capture both above laws, a shape of the plot of group centrality measure with respect to the group size should resemble the shape of unimodal functions, i.e. as the group S is growing, its value is increasing but, at some point it starts decreasing (see Preliminaries).

Law of unimodality. *A group centrality measure should have the property of unimodality.*

We believe that the Law of unimodality should naturally hold for most of real-world networks, whenever a reasonable group measure is used. Again, the above law is not meant to hold strictly and is therefore not stated in a mathematically precise form. Although the above mentioned laws seem very natural, we are not familiar with any group centrality mechanism to automatically deal with the groups of different sizes. It seems to us that the Freeman centralization approach [16] can be accustomed to achieve this, and is hence encapsluated in this paper. We focus on a particular type of group centralization, namely *group degree centralization*, as it is the simplest one to deal with.

Our experiments on various real-world networks shows how the values of group degree centralization change with respect to the group sizes. In the experiments, the law of unimodality is observed, which suggests that the Freeman centralization of degree may be a reasonable measure to consider while solving Q2.

6 Conclusions and future work

With respect to the important branches of the centrality theory mentioned in Introduction, we seek to identify the most 'influential' group of nodes in a network, without fixing the group size k in advance. While some of the related work regarding group centrality measures (i.e. question Q1), is already discussed in the Group centrality section above, there is (to our knowledge) no literature that would discuss the questions Q2 and Q3. The main problem is, that the values of two centrality measures may not be comparable, if the corresponding group cardinalities are not equal. Furthermore, while *Law of small groups* applies to most of known group centrality indices, this is not the case with *Law of unimodality*.

With regard to this, the Freeman centralization approach seems to be the only established method that efficiently normalize the result, so that the centrality scores can be compared in a meaningful way, regardless of their corresponding group cardinalities.

For further work we propose the following. While in the paper we only study the group degree centralization, one may study group centrality measures of some other centrality indices such as betweenness, closeness or eccentricity. In particular, it would be very interesting to verify whether for real-world networks the centralization variants of those proposed measures satisfy the above discussed laws, as well as the unimodality property. We study the Freeman centralization approach, as it seems the only established procedure of normalization, compatible with Q2. Alternatively, one could put aside the Freeman centralization and consider a different type of normalization for group centrality measures, that could preferably be more efficient to calculate.

References

[1] Bavelas, A.: A mathematical model for group structures. Hum. Organ. **7**(3), 16–30 (1948), `https://doi.org/10.17730/humo.7.3.f4033344851gl053`

[2] Bavelas, A.: Communication patterns in task-oriented groups. J. Acoust. Soc. Am. pp. 725–730 (1950), `https://doi.org/10.1121/1.1906679`

[3] Bell, J.R.: Subgroup centrality measures. Network Science **2**(02), 277–297 (2014), `https://doi.org/10.1017/nws.2014.15`

[4] Borgatti, S.P.: Identifying sets of key players in a social network. Comput. Math. Organ. Theory **12**(1), 21–34 (2006), `https://doi.org/10.1007/s10588-006-7084-x`

[5] Brandes, U., Erlebach, T.: Network analysis: methodological foundations, vol. 3418. Springer Science & Business Media (2005), `https://doi.org/10.1007/b106453`

[6] Estrada, E., Rodriguez-Velazquez, J.A.: Subgraph centrality in complex networks. Physical Review E **71**(5), 056103 (2005), `https://doi.org/10.1103/physreve.71.056103`

[7] Everett, M.G., Borgatti, S.P.: The centrality of groups and classes. J. Math. Sociol. **23**(3), 181–201 (1999), `https://doi.org/10.1080/0022250x.1999.9990219`

[8] Everett, M.G., Borgatti, S.P.: Extending centrality. Models and methods in social network analysis **35**(1), 57–76 (2005), `https://doi.org/10.1017/cbo9780511811395.004`

[9] Everett, M.G., Sinclair, P., Dankelmann, P.: Some centrality results new and old. J. Math. Sociol. **28**(4), 215–227 (2004), `http://doi.org/10.1080/00222500490516671`

[10] Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks **1**(3), 215–239 (1979), `https://doi.org/10.1016/0378-8733(78)90021-7`

[11] Hall, B.H., Jaffe, A.B., Trajtenberg, M.: The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. NBER Working Papers 8498, National Bureau of Economic Research, Inc (2001), `https://doi.org/10.3386/w8498`

[12] Hochbaum, D.S., Pathria, A.: Analysis of the greedy approach in problems of maximum k-coverage. Naval Research Logistics **45**(6), 615–627 (1998), https://doi.org/10.1002/(SICI)1520-6750(199809)45:6<615::AID-NAV5>3.0.CO;2-5

[13] Jordan, C.: Sur les assemblages de lignes. J. Reine Angew. Math **70**(185), 81 (1869), https://doi.org/10.1515/crll.1869.70.185

[14] Koschützki, D., Lehmann, K.A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O.: Centrality indices. In: Network analysis, pp. 16–61. Springer (2005), https://doi.org/10.1007/978-3-540-31955-9_3

[15] Krnc, M., Škrekovski, R.: Group degree centrality and centralization in networks. Mathematics **8**(10) (2020), https://doi.org/10.3390/math8101810

[16] Krnc, M., Škrekovski, R.: Group centralization of network indices. Discrete Appl. Math. **186**, 147–157 (2015), https://doi.org/10.1016/j.dam.2015.01.007

[17] Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web **1**(1) (2007), https://doi.org/10.1145/1134707.1134732

[18] Leskovec, J., Kleinberg, J., Faloutsos, C.: 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. pp. 177–187. KDD '05, ACM, New York, NY, USA (2005), https://doi.org/10.1145/1081870.1081893

[19] Leskovec, J., Mcauley, J.J.: Learning to discover social circles in ego networks. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 548–556 (2012), https://doi.org/10.1145/2556612

[20] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. In: Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07). San Diego, CA (2007), https://doi.org/10.1145/1298306.1298311

[21] Miyano, E., Ono, H.: Maximum domination problem. In: Proceedings of the Seventeenth Computing: The Australasian Theory Symposium-Volume 119. pp. 55–62. Australian Computer Society, Inc. (2011)

[22] Proctor, C.H., Loomis, C.P.: Research Methods in Social Relations, chap. Analysis of sociometric data, pp. 561–586. Dryden Press, New York (1951)

[23] Seeley, J.R.: The net of reciprocal influence; a problem in treating sociometric data. Can. J. Psychol. (3), 234–240 (1949), https://doi.org/10.1037/h0084096

[24] Vohra, R.V., Hall, N.G.: A probabilistic analysis of the maximal covering location problem. Discrete Applied Mathematics **43**(2), 175–183 (1993), https://doi.org/10.1016/0166-218x(93)90006-a

[25] White, S., Smyth, P.: Algorithms for estimating relative importance in networks. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 266–275. ACM (2003), https://doi.org/10.1145/956750.956782