universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Zentralitätsbegriffe in sozialen Netzwerken“ /

„Centrality in Social Networks“

verfasst von / submitted by

Johanna Spiegelhofer

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 915
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Betriebswirtschaft UG2002
Betreut von / Supervisor:	Univ.-Prof. Dr. Thomas Pfeiffer

Table of Contents

List of Figures

1. Introduction

Suppose you could choose any position within a given network. Which one would it be? Social Network Theory in its modern form allows the study of networks comprised of people, organizations or groups. It provides tools to make assumptions and predictions about connections between actors, their relationships, and the way goods travel between them. For example, *propinquity* suggests that individuals are more likely to be connected if they are geographically close (Kadushin, 2012; Feld & Carter, 1998), an aspect that is of importance when choosing a company's headquarters or making decisions about the location of research and development centres. The concept of *homophily* explains that actors with an uncommonly high number of matching characteristics are more likely to be connected (Lazarsfeld & Merton, 1954; Feld & Carter, 1998; Burt, 1982) – a quality that is of importance whenever teams are assembled and decisions about expected creativity and diversity need to be made. To some extent, social network theory can go as far as assessing an individual's importance within a network. Granovetter (1973) made the discovery that loose acquaintances, or *weak ties*, can be a source of valuable information because they allow access to remote network areas. Burt (2000) made similar findings about individuals connecting otherwise isolated areas of a network. Such results are important to identify key players within a network, whose value lies in the brokering of information or other goods. However, they cannot provide information about a randomly selected individual.

Graph Theory, a subfield of discrete mathematics, looks at networks from a purely analytical perspective. Stripping away all *social* aspects of a network, a structured frame of nodes (actors) and ties (connections) remains, which lends itself to abstract analysis. Sociology has taken advantage of this and harnessed graph theory for its use.

Regarding social networks from this reduced viewpoint introduces a plethora of evaluation methods that are based on the concept of *centrality*. Four of these are identified as important in literature and introduced in this work: degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality (Borgatti, 2005). Degree centrality bases its measurement on the number of direct neighbours and can – for example – help in identifying popular individuals in a group (Freeman, 1979). Such insight can be critical when information needs to be spread quickly among a wide number of people. Betweenness centrality tells us how often one individual lies on a shortest path connecting any pair of others (Freeman, 1977). This can be of importance, if the objective is to identify players who have the power to delay or interrupt information flow. Closeness centrality measures how close an actor is to all other actors combined (Bavelas, 1950). An important application of this measure lies in its ability to identify

influencers in a network, by recognising positions suitable for broadcasting. Lastly, eigenvector centrality gages an actor's importance by looking at its immediate neighbours (like degree centrality) but goes further and calculates those neighbours' importance before adding their values (Bonacich, 1972). This measure of centrality is at the core of the ordering of results in the search engine Google (Brin & Page, 1998) but can be applied anywhere, if the importance of one individual is based on the importance of its neighbours.

Based on these measures of centrality, graph theory provides the tools to evaluate network positions and assign a ranking in accordance with the derived results. However, not every method applicable to a graph yields reasonable results if it is used on a real-world problem. The betweenness centrality measure, for instance, deals with shortest paths. Applying it to a network where an arbitrary route is taken, like a drunkard finding his way home, can lead to misleading or plain wrong results (Borgatti, 2005).

In this work, we[1] will briefly recapitulate the underlying theories of social network theory, graph theory and tournament theory. We will proceed by presenting in detail the four types of centrality mentioned above. This should provide the reader with the necessary knowledge to understand how each measure is obtained, how to derive a ranking from it as well as how and when to use it. Also, various economic applications are mentioned to illustrate directly possible uses of the respective concepts. We will conclude by providing an in-depth discussion of scientific articles employing one of each centrality measures. This literature was chosen from a variety of areas to demonstrate versatility, reach and wide applicability of centrality analysis in social networks.

2. Materials and Methods

The theoretical aspects of this work were gathered through literature search in peer-reviewed journals, using internet search tools of the library of the University of Vienna. Hardcopies of subject literature were used to study selected elements in depth.

Simulations of Radicchi's (2011) diffusion algorithm using a variant of eigenvector centrality and creation of corresponding calculations and graphs as well as tables and graphs to illustrate aspects of closeness centrality and betweenness centrality were done using Microsoft Excel Version 1807.

All other illustrations were composed using Microsoft PowerPoint, Version 1808.

[1] Note that the use of "we" in this work does not indicate multiple authors but serves as an unpersonalized way to reference the author.

3. Underlying Theories

In the following section three research areas are introduced: social network theory, graph theory and tournament theory.

3.1 Social Network Theory

Modern Social Network Theory is mostly based on developments in different fields of science, taking place in the first half of the 20th century, mainly psychology, anthropology and mathematics. Psychologist and inventor of sociometry, Jacob Moreno (1889 – 1974) started to actively plot networks of people's circles of friends, the nature of their relationships and how these relationships translated into benefits and limitations. By mapping these networks, using points to represent people and lines to represent relationships, he invented the sociogram, the now common way to graphically represent social links. This facilitated the identification of social patterns, such as chains of connections, "stars", or "isolates". At roughly the same time, Kurt Lewin (1890 – 1947), also a psychologist, applied mathematics to study group behaviour, using topology and linkages through vectors also by means of points and lines, to create networks. In the field of anthropology, Alfred Radcliffe-Brown (1881 – 1955) and Siegfried Nadel (1903 – 1956) realized that social structures (and thus societies) are subject to individual relationships, forming a connected web – or network. Mathematics provided a language to analyze and measure social network phenomena, translating real-life networks into abstract patterns (Scott, 1997).

The famous "Hawthorne studies" (1924 – 1932), commissioned by the US American National Research Council, with pioneering work done by Elton Mayo (1880 – 1949) and William Warner (1898 – 1970), also fell into this timeframe. They uncovered the now well known "Hawthorne effect" or "observer effect" in social systems, which describes that the knowledge of being observed leads individuals to adjust their behaviour, leading (among other things) to a temporary increase in productivity (Scott, 1997; McCarney, et al., 2007).

This implied not only that observation can change people's behaviour, but work performance and productivity are linked to social structures besides objective measures, such as pay (Scott, 1997; McCarney, et al., 2007).

These (and other) pioneers laid the groundwork for Mark Granovetter (*1943) and others to delve into complex social network analysis as we know it today.

Since social network analysis originates in a number of different fields of science, there is often more than one definition for any of the formal terms used.

3.1.1 What is a Network

Kadushin (2012) chooses a mathematical approach: *"A network is a set of relationships. More formally, a network contains a set of objects (in mathematical terms, nodes) and a mapping or description of relations between the objects or nodes"* (Kadushin, 2012, p. 14).

Mitchell (1969) defines networks in terms of social sciences, defining nodes only as people:

> A network is *"(...) a specific set of linkages among a defined set of persons with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the persons involved"* (Mitchell, 1969, p. 2).

Wasserman and Faust (2008) emphasize the importance of relations between nodes and their necessary connectedness:

> A social network *"...consists of a finite set or sets of actors and the relation or relations defined on them. The presence of relational information is a critical and defining feature of a social network." "A social network arises when all actors can, theoretically, have ties to all relevant actors"* (Wasserman & Faust, 1994, p. 20).

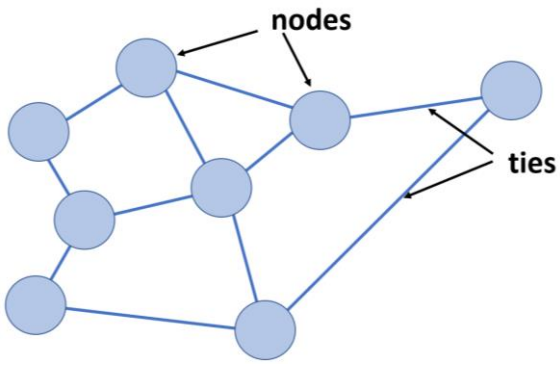


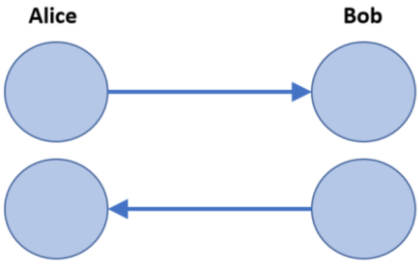
Figure 1. A network consisting of eight nodes and ten ties.

In its most basic setup, all that is required to plot a network that can be formally analyzed, are a pair of objects (often referred to as "nodes") and a (definite) relationship between them, represented through a connecting edge (or tie). These properties are sufficient to create a network that can be mapped and analyzed (Kadushin, 2012). An example for such a minimum network is a pair of friends.

Even this basic setup allows for three significantly different scenarios, depending on the relationship between nodes. Referring to object 1 as A (Alice) and object 2 as B (Bob), the following relationships, as described by Kadushin (2012), can be considered:

1. the relationship between A and B is **undirected**. This can be visualized as a simple edge between A and B. A practical example for such a relationship would be Bob and Alice having the same mother.

2. the relationship between A and B is **directional**. This can be visualized as an arrow between A and B. Either in the direction $A \rightarrow B$ or $B \rightarrow A$. Examples for such relationships could be Alice being the daughter of Bob or Bob being the son of Alice. Both could not be true in both directions. Consequently, it is not irrelevant in which direction the arrow's head is directed.

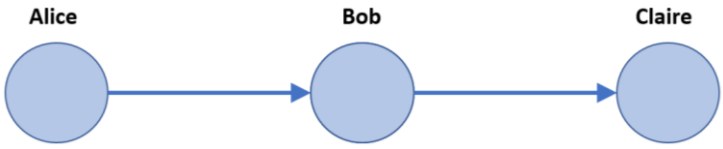


3. the relationship between A and B is **symmetrical**. In this case, the relationship is true in both directions. An example for such a relationship is Alice being married to Bob and Bob being married to Alice, with the connecting arrow having heads on both ends to visualize the mutual expression.

At this point we emphasize that all immediate connections described are between exactly two nodes. (We note that the mathematical concept of hyper-graphs, which is part of graph theory and allows for more types of connections between nodes, is not considered in this work.) This however does not prohibit more than one edge leaving each node.

Increasing the scope of the network by admitting additional nodes, allows for further scenarios, as all of the above relationships (undirected, directional, symmetrical) are possible between

each connected pair of nodes. But there is also the possibility for a new "role": Adding one additional node C (Claire), allows for the role of intermediary. A might be connected to B and B connected to C. Even though there is no immediate connection between A and C a path between them can be expressed, using B as an intermediary.



"*Where three elements A, B, C, constitute a group, there is, in addition to the direct relationship between A and B, for instance, their indirect one, which is derived from their common relation to C",* (Simmel, 1950).

Triads therefore allow for the possibility of an indirect relationship, which is impossible in a dyad, where there can only either be a connection – or not. The additional object can adopt roles impossible in a dyad: he/she might align with one of the others, gaining an advantage over the third, he/she might act as a mediator or broker, gaining individual advantage in both cases or be neutral. These new options can increase the complexity of a dyad network by far (Kadushin, 2012).

Another possibility is for all three nodes to be mutually connected, creating a triangle – and in certain cases transitivity. Such groups of three (or triads) are the smallest social structures studied in social network theory that display the characteristics of a society, and where real hierarchical development can be observed (Kadushin, 2012; Hanneman & Riddle, 2005).

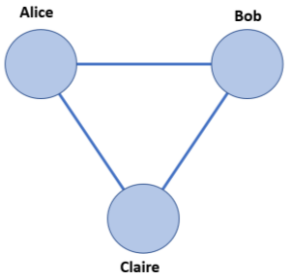


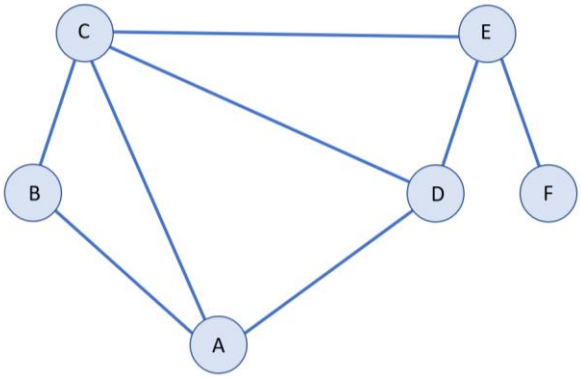
Figure 2. A triad.

As indicated above, the number of possible network structures rises significantly with the number of nodes (Kadushin, 2012).

3.1.1.1 Visualizing networks

There are two common ways to depict social networks: sociograms and matrices. The term *sociogram* was introduced by Moreno and is used for a visual diagram of social networks showing nodes as dots and connections as lines. This allows for easy peripheral understanding of networks as clusters or areas of high or low density that can be recognized at a glance and are intuitive to understand (Kadushin, 2012). Figure 1 above can be interpreted as a sociogram.

For complex network analysis, matrices are often preferred. *Adjacency matrices* are sociograms displayed in a more mathematically formal way, with the rows and columns of the matrix indicating nodes and the values in the matrix indicating the presence (or absence) of a tie (Kadushin, 2012). *Figure 3* shows two representations of an identical network, once depicted as a sociogram and once by ways of an adjacency matrix.



	A	B	C	D	E	F
A	0	1	1	1	0	0
B	1	0	1	0	0	0
C	1	1	0	1	1	0
D	1	0	1	0	1	0
E	0	0	1	1	0	1
F	0	0	0	0	1	0

Figure 3. Sociogram (left) and corresponding adjacency matrix (right) of an undirected network with six nodes and eight ties.

Where structurally equivalent nodes form clusters, these can be represented by "blocks" in an adjacency matrix, forming a *blockmodel* (Burt, 1992).

	A	B	C	D
A	0	0	1	1
B	0	0	1	1
C	1	1	0	0
D	1	1	0	0

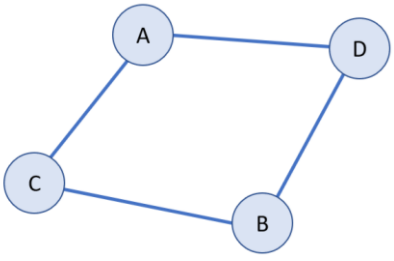


Figure 4. Sociogram and blockmodel of an undirected graph with four nodes.

3.1.1.2 Transitivity

This concept should not be confused with *flow*. While "flow" is established whenever a path exists from any start to any end-point, *transitivity* has a formal mathematical definition:

Whenever there exists a path of length 2 (from A to B and from B to C), there must also be a direct path from start- to end-point (A to C):

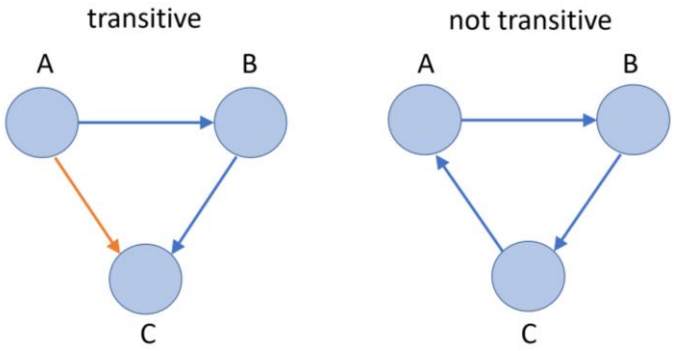


Figure 5. A transitive and a not-transitive triad. The second triad cannot be transitive because there is the need for a "broker" between non-consecutive nodes.

In bigger networks, with paths of any length, the network is said to be transitive, if all paths of length 2 fulfil the above definition.

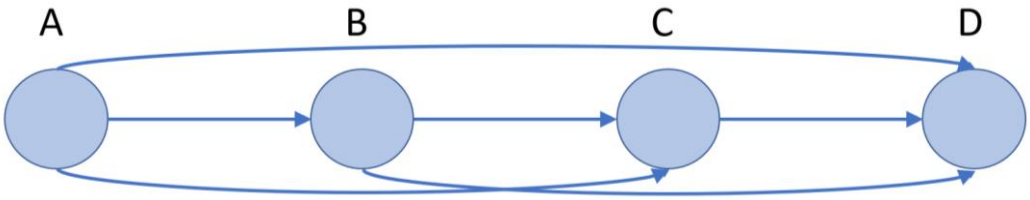


Figure 6. Transitive relationship between four nodes A, B, C and D.

Transitivity is an important feature in social networks, as these display a natural tendency to form clusters. A common example for this development is a group of friends, whose members tend to develop friendly relationships as well (Holland & Leinhardt, 1971; Davis, 1970). This leads to transitive sub-graphs. Another example for a transitive relationship could be superiors: If Alice is Bob's superior and Bob is Claire's superior, then Alice is also Claire's superior.

In some instances, transitivity is not a desirable network structure, as it can lead to a big number of redundant contacts, meaning that they are prone to having access to the same or similar resources.

3.1.2 Network structure

In terms of network structure, Kadushin (2012) defines three distinct set-ups that are commonly studied: *Ego-centric*, *socio-centric* and *open-system* networks. The first two are currently the main approaches used in social network studies (Chung, Hossain, & Davis, 2005).

Ego-centric: in this setup all nodes within the network are connected to one central node. There may also be connections between other nodes, but the focus lies on one main individual instead of the network as a whole (Chung, Hossain, & Davis, 2005). As noted above, it is an essential premise for network analysis, to have information about the relationships between nodes. If information is, for example, obtained through questionnaires, the resulting data might only allow for the establishment of ego-centric networks, as connections between other than the questioned individuals, cannot be identified (Kadushin, 2012; Hanneman & Riddle, 2005).

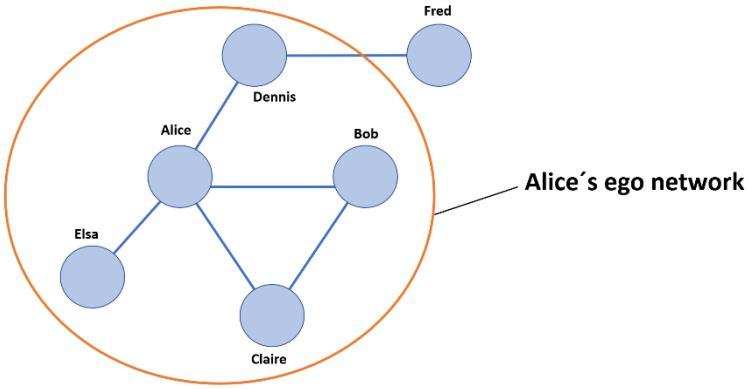


Figure 7. Ego network with Alice at its centre. Fred is not directly connected to Alice and therefore not part of this network.

From any network, each node's ego-network can be extracted, resulting in many individual ego-networks (Hanneman & Riddle, 2005). *Figure 7* shows Alice's ego-network, which does not contain Fred since there is no immediate connection between the two. This network is extracted from a bigger network that includes Fred.

An example for an ego-centric network would be Alice's set of friends. They might or might not be mutually acquainted, but they are all connected to Alice. Fred on the other hand is not part of Alice's set of immediate friends and therefore not part of her ego-network.

Socio-centric: this set-up does not require one central node, instead it focuses on the network as a whole (Kadushin, 2012; Chung, Hossain, & Davis, 2005). Bernard et al. (1989) refer to socio-centric networks as "networks in a box", meaning that the network consists of a predefined group. Examples for socio-centric networks would be the network of every tennis player in the ATP world tour of a certain year or the network of every student enrolled in a certain university in a given term (Kadushin, 2012).

In both cases the networks are closed in terms of boundaries and there is no uncertainty pertaining as to who is part of the network. Socio-centric networks are most commonly studied if the available data allow for their construction (Kadushin, 2012).

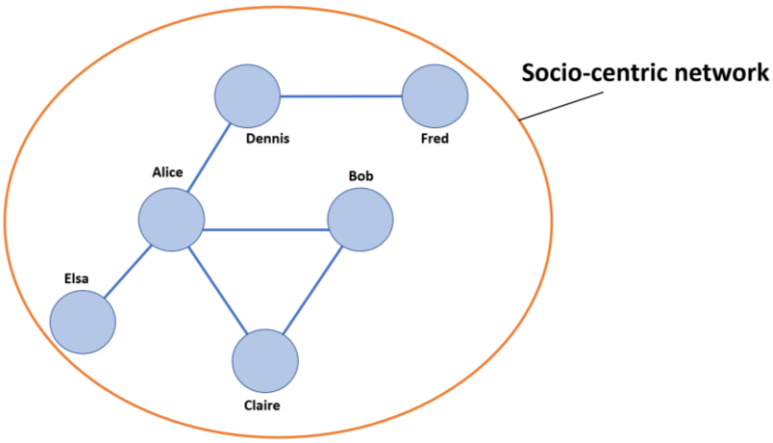


Figure 8. Socio centric network.

Open system: open system networks differ from socio-centric networks in that their boundaries are either not set at the beginning or cannot be known (without unreasonable effort). An example would be all people interested in music. While theoretically every person in the world could conceivably be questioned, the actual prospect of doing so is unfeasible, establishing the relationships between these nodes even more so (Kadushin, 2012). For analytical reasons, open systems are more difficult to study: "*The restriction to a finite set or sets of actors is an analytic requirement. Though one could conceive of ties extending among actors in a nearly infinite group of acts, one would have great difficulty analyzing data such a network*" (Wasserman & Faust, 1994, p. 19).

Figure 9 shows an example of an open system network. For obvious reasons an infinite number of nodes cannot be portrayed.

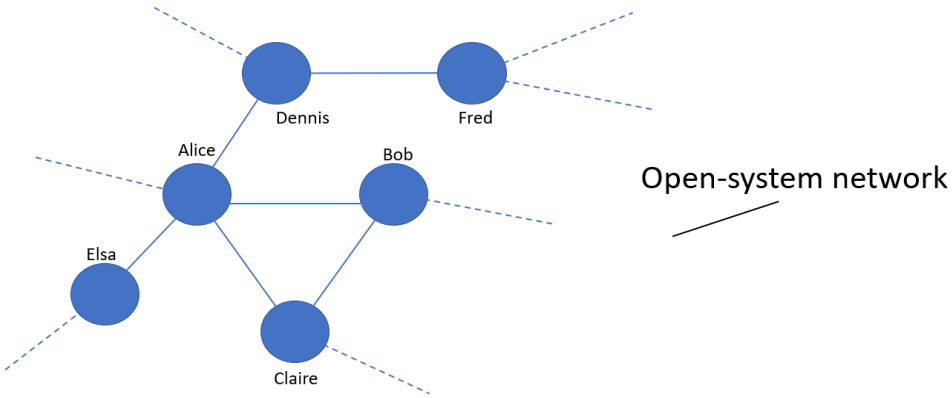


Figure 9. Open system network. Dotted lines represent ties to additional nodes to indicate the absence of a boundary.

Selecting the boundaries of a network can be challenging but it is a necessary requirement for meaningful network analysis. Both actors and the nature of their ties needs to be defined

(Kadushin, 2012). "*The boundary of a set of actors allows a researcher to describe and identify the population under study"* (Wasserman & Faust, 1994, p. 31).

All three types of networks allow for the analysis of networks of different orders of magnitude. From the simple network of a group of friends to the complex relationships between international organizations (Kadushin, 2012).

3.1.2.1 Network connections

There are a number of factors that determine the likelihood of two nodes being connected within a social network. Two important factors are discussed below:

Propinquity: Geographical closeness increases the likelihood of a pair of nodes being connected. This holds true for different kinds of networks, from housing projects where neighbours are more likely to be friends (Kadushin, 2012; Feld & Carter, 1998), to people serving on multiple boards of directors if firms' headquarters are in close vicinity (Kono, Palmer, Friedland, & Zafonte, 1998)

Homophily: If a pair of nodes has more matching characteristics than would be expected from a random sample in their network, this is considered homophilous and increases the likelihood of this pair to be connected (Lazarsfeld & Merton, 1954). However, there are two possible sources for homophily, as defined by Burt (1982): 1) common attributes in connected nodes may lead to common norms and 2) common norms may lead to a connection between nodes with common attributes. Feld and Carter (1998) on the other hand also consider a pair of nodes' "structural location" within networks to be a source for homophily. A distinction between homophily and propinquity is not always intuitive as both might occur simultaneously or might even cause one another (Vernon & Hoover, 1962).

3.1.2.2 Social Capital

Social Science focuses on "*social capital*", a concept of value derived from a network or resources within an individual's accessible network. This form of capital is often considered a measurable resource, like economic capital, though it is wrought with difficulty to establish rules to derive a monetary-equivalent measure of its value (Bourdieu, 2011).

In any case, social capital cannot exist in isolation, but only within a group: "*Unlike other forms of capital, social capital inhers in the structure of relations between actors and among actors"* (Coleman, 1988, p. 98). Burt (2000) also determines that social capital is dependent on an actor's ability to draw profit from their position within a network. There, relations can lead to profit chances for economic or social capital but are also determined by network structure and

each individual's position within that structure (Kadushin, 2012; Coleman, 1988; Bourdieu, 2011).

3.1.3 Network Analysis

Originating from diverse fields of research, network analysis is met with different limitations but also requirements and focus areas, depending on the field it is applied to.
Social Science for instance faces limitations where certain network conditions cannot be achieved or are exceedingly unlikely to occur. Social structures (like hierarchy, social circles, etc.) can prevent ties between different levels and limit network connections. Other social aspects (members of direct neighborhood, group of co-workers, etc.) also have the tendency to shape networks, limiting possible structures and network shapes (Kadushin, 2012).

3.1.3.1 Strength of weak ties

Mark Granovetter (1973) coined the now famous concept of the "strength of weak ties". He found that there is a tendency to form dense groups, or clusters, among strong ties (ties that form a strong connection like the circle of close friends), that tend to be connected through weak ties (ties that are only loosely connected, like acquaintances). *"... [Our] acquaintances ("weak ties") are less likely to be socially involved with one another than are our close friends ("strong ties"). Thus, the set of people make up on any individual and his or her acquaintances will constitute a low-density network (one in which many of the possible ties are absent), whereas the set consisting of the same individual and his or her close friends will be densely knit (many of the possible lines present)* (Granovetter, 1982, p. 105).
One of Granovetter's key ideas is that weak ties allow access to remote parts of a network because they are more likely to be connected to a different set of nodes than the ego. In social relations this implies that gaining access to "new" people (and the corresponding resources) is more likely to occur through an acquaintance than it is through a close friend because it is likely you already know your friends' friends: *"... Ego will have a collection of close friends, most of whom are in touch with one another – a dense "clump" of social structure. Ego will have a collection of acquaintances, few of whom know one another. Each of these acquaintances, however, is likely to have close friends in his or her own right and therefore to be enmeshed in a closely-knit clump of social structure, but one different from Ego's",* (Granovetter, 1982, pp. 105).
These weak ties allow the formation of "bridges" between clusters, connecting otherwise remote or possibly inaccessible areas and "opening" closed networks. This leads to better dispersion of knowledge within a network and can serve to increase overall creativity since ideas can

travel “far and wide”. Access to different clusters is highly valuable for resource flow and for limiting redundancy – therein lies the “strength” of weak ties. Network structure, as it is depicted using sociograms, can help to detect valuable bridges (Granovetter, 1973). *Figure 10* shows an example of a network consisting of two clusters $A-B-C-D-E$ and $F-G-H-I-J$ that are bridged only by the tie $E-F$. Absence of this crucial tie would entail a break up of the big network into two separate networks.

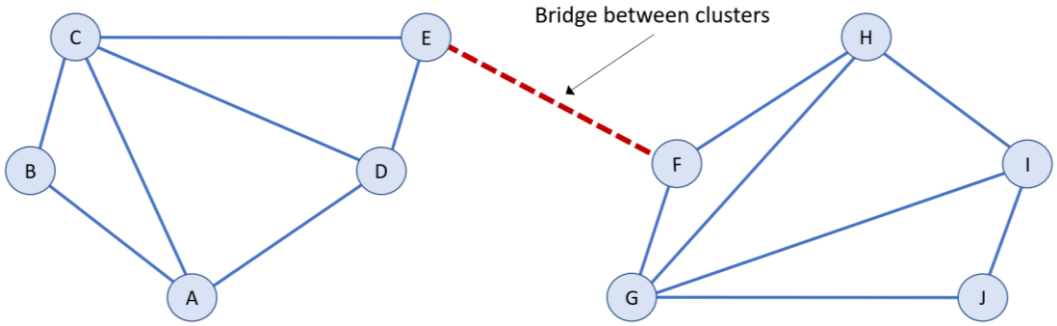


Figure 10. A network of ten nodes with the bridge E-F serving as the only connection between two clusters.

However, it is rarely the case in real life that individual clusters are totally isolated and only connected through single bridges. Instead, there tends to be an overlap of groups to some degree and sometimes even strong ties can be responsible for the formation of a bridge (Kadushin, 2012).

Both strong and weak ties have positive and negative connotations. As explained above, weak ties facilitate access to distant areas of a network and thus novel information, but strong ties allow for the development of trust (Kadushin, 2012; Granovetter, 1982). This trust will prevent actors from behaving self-interestedly, improving collaboration, knowledge diffusion and member contributions – important factors for team activities (Levin & Cross, 2004). In case trust is violated, social sanctions can be applied, the knowledge of this works as a deterrent factor (Kadushin, 2012). Apart from that, costly transactions are more likely to be fulfilled through strong ties because nodes are often unwilling to spend much of their resources on a flimsy connection. It is important to note that, while weak ties are more frequently bridges than strong ties, most weak ties are not bridges (Kadushin, 2012; Granovetter, 1982).

3.1.3.2 Distance

The number of “steps” or edges between nodes is often referred to as “distance” between these within a network (Kadushin, 2012). In many cases, however, there are multiple paths connecting any pair of nodes, creating the question of which path to select when determining the gap between them. Often, the length of a shortest path is used to determine the distance.

Another definition of distance uses the number of redundant paths to determine closeness: the greater the number of paths, the "closer" a pair of nodes. Generally speaking, different analytical techniques use different definitions of distance (Kadushin, 2012). Considering this, it is vastly important to know what an author understands under the notion of "distance" to properly understand his results.

Stanley Milgram (Travers & Milgram, 1967) raised the subject of distance between people (distance in this case being the number of steps, in terms of personal contacts, separating them), when he published his paper "The Small World Problem" in 1967 and subsequent studies. In these studies, he found that the estimated number of steps between any pair of two people in the United States of America is six, meaning that on average, five intermediate people are necessary to connect any pair of two (Travers & Milgram, 1967). Several interesting phenomena were discovered during this research. For one, links between men and women were discovered to be less frequent than between the same sex, also social barriers were proven to be difficult to overcome.

3.1.3.3 Network flow

If relationships along a given graph are either symmetrical or directional, a transfer of resources, so called "*flow*", can be established along its edges. The transferred resources can be either material or non-material, consisting of actual tangible goods, information, contracts or suchlike (Borgatti, 2005; Kadushin, 2012; Wasserman & Faust, 1994). For example, if A has access to information of B and B has access to information of C, then A can gain access to C's information, allowing for a "flow" through the intermediary B. Depending on an individual's position within a given network, his/ her social capital may be higher or lower, depending access to resources and thus (Kadushin, 2012).

Flow through networks can be classified further, depending on the nature of goods (divisible or indivisible) that travel and the route that is selected through the network. Borgatti (2005) classifies network flow according to these two parameters. He distinguishes four kinds of routes (or *trajectories*) that goods can follow:

- *walk*: an unrestricted route between A and B.

  The route $A - D - E - B - A - D - B$ is an example of a walk, where nodes A, B and D are visited multiple times, as is the edge $A - D$. Shared coffee cups in an office or the sequence of positions of a knight on a chessboard during a game give examples of typical walks.

- *trail*: a route between A and B, where no edge is visited twice.

  The route $A - D - E - B - A - E$ is an example of a trail, where nodes A and E are visited multiple times, but no edge is used more than once. A recipe passed around within a group of friends, for example, may reach the same person more than once, but it is unlikely that this would happen through the same person.

  Note that every trail is a walk.

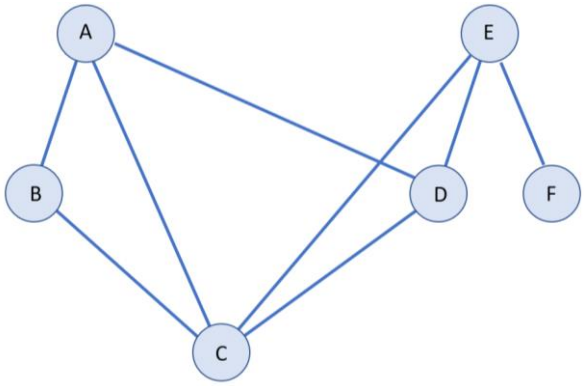
- *path*: a route between A and B, where no node is visited twice.

  The route $A - B - C - D - E$ is a path. Neither nodes nor edges are travelled multiple times. A birthday card (in the optimal case) will travel along a path, since everybody who signed the card passes it on to somebody who has yet to do so.

  Note that every path is a trail.

- *geodesic*: shortest route between two nodes A and B, which means that the smallest possible number of nodes between A and B is visited. (The length of an edge in a graph is purely dependent on purposes of visualization and has no significance for the graph-theoretical length of a route.)

  In the graph below, both routes $A - B - E - F$ and $A - C - E - F$ connecting A and F are examples of geodesics. Note that every geodesic is a path.



Further, Borgatti (2005) suggests that it is important to differentiate between three methods of spread of goods: broadcast, serial replication, and transfer. The possible method of spread of any good depends on its characteristics: divisible or indivisible.

- *transfer*: move mechanism for indivisible goods. Goods that cannot be divided have the property that they can only be passed on serially. They cannot be in two places at one time and therefore have to be passed on from node to node, like a football being passed between players (nodes) during a match or the office stapling machine.

- *serial duplication*: copy mechanism for divisible goods from one node to one other node. Under "divisible" we understand both actual divisibility (like water poured into two glasses) and the possibility of duplication (the stapler above cannot be duplicated, whereas a catchy tune can easily be copied to another person, with the result that both people suffer it).
- *parallel duplication* (*broadcast*): copy mechanism for divisible goods from one node to at least one other node. An example for this mode of travel is dissemination of knowledge through a presentation held by one person to an audience of many.

In network analysis, it is important to understand which "mode" of travel is used by which kind of good. Employing incompatible concepts and measures can lead to distorted or simply wrong results (Borgatti, 2005).

3.1.3.4 Position - Centrality

As discussed above, connections between nodes can be directional or un-directional, with flow possible between relationships and a vast increase of network complexity for networks containing three or more nodes. Also discussed is the importance of roles a "third" node in a triad can adopt. Keeping this in mind, and looking at more complex networks, the concept of position within a network needs to be discussed. It is not only important how "far" a pair of nodes is apart, but also how each node is positioned within a network (central or on the periphery, in a dense or sparsely occupied area or anywhere in-between). Any "location" has distinct characteristics that can be described using different types of the concept of "centrality". Centrality in its most basic idea, describes how close to the center a node is located within a network (Kadushin, 2012).

Structural similarity implies that comparable network patterns or building blocks bring forth similar roles: "*Nodes that have similar patterns of relationships with other nodes are* […] *grouped together,"* (Burt, 1992). At the same instance, a social network is a flexible structure that is prone to change whenever any of its nodes changes its position, set of acquaintances or the like. A studied network is therefore only ever a snapshot in time of a network at a certain instance in time.

In social networks, centrality can relate to concepts like hierarchy: "*Hierarchy in networks is stated strictly in terms of location of a given node relative to other nodes, without assuming any content to this position"* (Kadushin, 2012, p. 40). Some nodes will be important simply because of their position and ability to connect individual groups (see also "strength of weak ties" above)

or facilitate connection between them (Burt, 2000; Kadushin, 2012). However, all positions in a network are necessarily defined by the nature of their connection (supervisor-employee; friends; kinship; …), leading to certain structural patterns occurring repeatedly and in independent networks. In real-life scenarios, relationships between people are complex, multilayered structures and individuals are connected through more than one set of characteristics. Such social networks are made up of a number of overlapping networks with a multitude of different types of connections between them. Separating which network is responsible for any single flow can be impossible. This overlapping of linking properties is called "multiplexy" but beyond the scope of this work, where each network is connected only through one kind of tie (Kadushin, 2012).

Different measures of centrality and their implications and applications are discussed below.

3.2 Graph Theory

As part of discrete mathematics, "graph theory" is concerned with abstract objects called *graphs* that consist of *nodes* and *ties* (corresponding to actors and their connecting characteristic in social networks) and the structure and ordering they enclose. A *directed graph* $G = (V, E)$ is defined as an ordered pair consisting of a set V of vertices (or nodes in network theory) and a set E of edges (or ties in network theory). Each element of E is an ordered pair of elements of V. Such an element (v, w) can be visualized as an arrow starting at v and ending in w. *Undirected graphs* $G = (V, E)$ are defined similarly but with the difference that E consists of sets $\{v, w\}$ of two elements, representing an undirected edge that can be visualized as a connecting line (Chartrand, 1984).

A generalization of graphs are *weighted graphs*. These are directed or undirected graphs with weights attached to the edges. Formally this can be realized by assigning a function f to the graph $G = (V, E, f)$, which ascribes a real number to each edge e. Alternatively, we can choose a function w_{ij} that allocates a weight to each pair of nodes (in this way we can dispense with the set E which could be recovered as the set of pairs having non-zero weight) (Chartrand, 1984). We will meet this concept in the chapter on eigenvector centrality as in the subsequent simulation part.

Social Network Analysis employs methods developed in graph theory to model, study and evaluate social relationships, which are "translated" into networks. Networks are graphs with additional information about nodes and edges, such as names or ages of individual people in a group of friends or the nature of their connection in a social network. Analysis of such networks can be done with the use of software solutions (Butts, 2008)

3.2.1 Ramsey's Theorem

An interesting mathematical theorem originating in graph theory is called "Ramsey's theorem" and deals with the inevitability of certain structural properties that are bound to arise in graphs. Each of these properties will occur in a graph of sufficient size. More specifically, given any number k, each sufficiently large undirected graph contains k nodes that either form a complete graph or the graph with no connections at all (Ramsey, 1930). A popular example for Ramsey's theorem, that has entered mathematical folklore, is the case $k = 3$, commonly known as the "theorem on friends and strangers".

Claim: Consider 6 people in a room. If any pair is either acquainted or not (there is an edge between them or not), there is at least one cluster of three that either know each other or do not.

Proof: The proof for this special case is intriguing in its simplicity. Leader (2001) uses a network of color coded edges: consider a node A (Alice), at a party of at least six people. There she will either know (red edge) three people or be a stranger (blue edge) to three people. Consider without loss of generality that she knows three people B (Bob), C (Claire) and D (Dennis), the other case would be analogous, swapping red and blue. One of two cases will occur. First case: there are two out of B, C and D that know each other. In this case Alice and these two people know each other and form a monochromatic (red) triangle (*figure 11*):

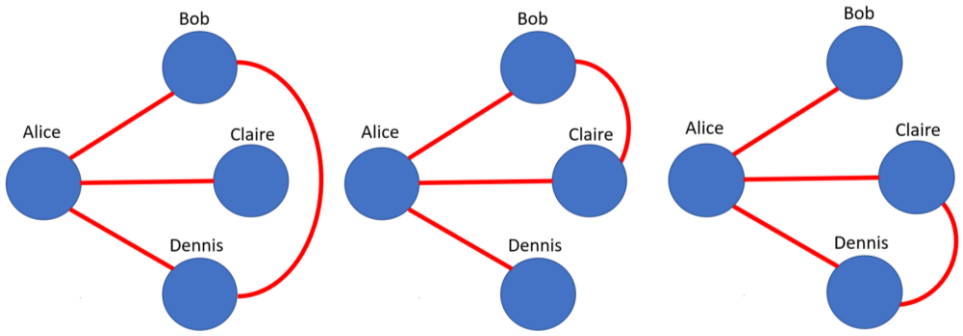


Figure 11. Ramsey's Theorem for k=3 with Alice knowing three people. First case scenario: two of the people Alice knows are mutually acquainted. Red triangles indicate the groups of three A-B-C, A-B-D and A-C-D that can occur.

The other possible case is that Bob, Claire, and Dennis are mutual strangers. In this case, these three form a blue triangle, which finishes the proof of the case $k = 3$ of Ramsey's theorem, showing in particular that a graph of six nodes is sufficient. ■

Figure 12 depicts this possible scenario:

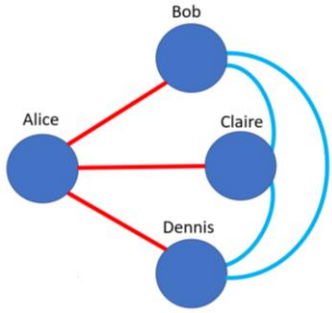


Figure 12. Ramsey's Theorem for k=3 with Alice knowing three people. Second case scenario: none of the people Alice knows are mutually acquainted. The blue triangle B-C-D indicates the group of unacquainted people that are all connected to Alice.

For a general k, the smallest size m of the graph needed in order to find a monochromatic complete graph of k nodes is not known. However, Ramsey showed that there is such an m.

3.3 Tournament Theory

We have discussed social network theory and graph theory in the previous sections. The third theory of interest to this work is *tournament theory* and more specifically, *rank-order-tournaments*.

Three kinds of tournaments are often studied:

- Optimal labor contracts are the subject of economic considerations,
- Abstract mathematical objects are studied in graph theory (see chapters on graph theory above and degree centrality below), and
- Sport-related tournaments we come across in everyday sporting events (see chapter on eigenvector centrality below).

This shows that the term tournament theory is used in different contexts and that the general idea of inducing competition to find an optimal outcome is of interest in diverse areas of science and everyday life in general.

The tournaments we consider in this work are such that their associated networks are relevant in real-life scenarios but can be modelled mathematically.

3.3.1 Background

In 1981, economists Edward Lazear and Sherwin Rosen developed the theory that optimal labor contracts might be evoked, if compensation was not based on absolute performance, but on relative performance instead (Lazear & Rosen, 1981). An important incentive for employers to choose such a compensation and selection system is that it attempts to circumvent the problem of finding absolute performance measures that would be costly and difficult to implement. As long as it is possible to rank performance, the incentive factor of a tournament should elicit high performance from employees.

Looking at this setting from a different perspective, it is of interest how any position within a tournament network is evaluated. Consider a network of international companies and a manager, who wants to take an influential position within this network by becoming the CEO of the company with the greatest centrality. Classic tournament theory deals with the question of how to obtain the sought position. Our view point on the other hand is how to assess the strength of each company's position by evaluating their respective measure of centrality: In this work the focus lies on achieving advantageous network positions.

3.3.2 Approach

We are not interested in absolute performance measures but in the ensuing ranking. In some cases, achieving a top rank may even be sufficient motivation to maximize effort (subject to utility) without an actual (monetary) prize being offered.

Let us consider the ranking of scientific journals, where a top rank is associated with high quality articles. Achieving a position within the network of peer-reviewed journals that improves its perceived public influence, is likely to increase such a journals dissemination. This in turn would increase the publications income and attract quality authors. Again, this increase in high performing authors would improve the ranking and the circle would start anew. Obviously, as there is a network of competing journals, they all have the same goal of achieving a highest possible rank. In such a model, ranking could result from the number of citations per article a journal receives. Below, such a network is discussed.

Consider a network made up of journals in the role of nodes and citations in the role of edges. Because scientists cite articles according to their relevance to their own work, they are likely to reference different journals, thus creating (and constantly changing) the network. *Figure 13* shows such a network of journals and the referenced articles that connect them.

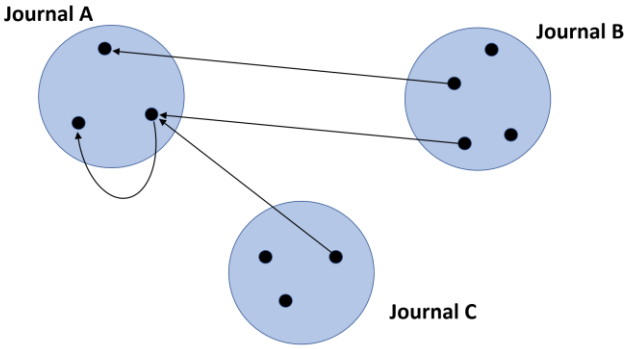


Figure 13. Arrows represent citations of articles in journal A by articles in journals A, B and C.

An edge from journal A to journal B is weighted according to the number of citations from articles in A to articles in B (figure 14). To obtain the necessary weights, these numbers of citations are divided by the number of articles published in journal A. By adding the weights of

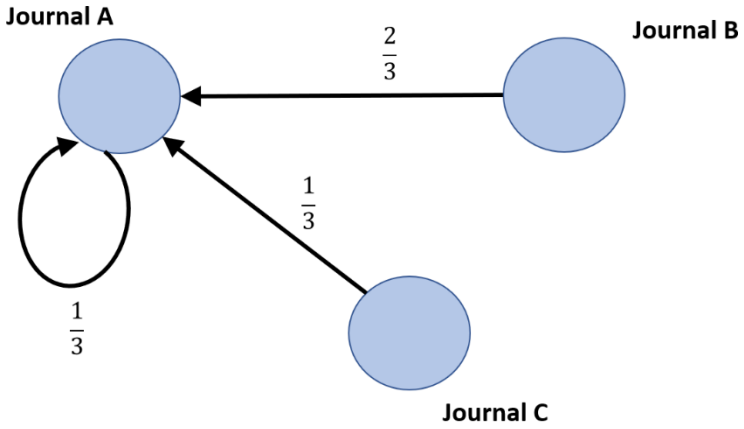


Figure 14. Weighted arrows indicate the ratio of citations by published articles.

incoming edges, an evaluation can be obtained. If this is repeated for all journals, a ranking of these journals is obtained. This system is used in the real world. The online platform MathSciNet (AmericanMathematicalSociety, 2018a) uses a very similar ranking method for their content. This webpage (with restricted access) is hosted by the American Mathematical Society (AMS) and contains mathematical reviews for a great number of research papers. It ranks journals according to a five year "mathematical citation quotient" or in short "5-year MCQ" (AmericanMathematicalSociety, 2018b) to provide users with information about the quality of the respective journals. Clearly, a journal with a higher MCQ is likely to contain articles of higher quality (on average). This measure of quality uses a limited timeframe for its assessment.

In short it works as follows: for any given year Y, citations by articles published in Y of articles in journal A that are no older than five years are summed up and divided by the number of

articles published in this journal in the same five-year period. This time constraint ensures that outlier years with a very high or very low number of citations are mediated and the measure is relatively stable. It also makes sure that singularly successful articles do not continually distort a journal's measure (Lazear & Rosen's article on rank-order-tournaments is cited regularly even today, the 5-year MCQ however, would not take it into account). It is important to keep in mind that only comparable journals (or whatever else is reviewed) should be ranked at the same time when this method is used. Failing to take similarity into account would result in misleading results as the number of both citations and articles greatly depends on issues like the popularity of a field of research or the number of scientists working in it.

Below, the chapter on degree centrality deals with such a ranking method: the *points system* and the corresponding generalization to weighted graphs.

We consider the striving towards a high ranking as a tournament, where the achieved rank is considered the prize and trying to achieve it elicits effort and implies motivation. A major difference to traditional rank-order tournament theory is that the prize-spread cannot be defined in a useful manner, considering it 1 between every pair of adjacent ranks would be possible but lacking significance.

Graph theory has a separate definition of the term "tournament". In this field of discrete mathematics, every directed graph where there is exactly one arrow (with one arrow-head) between each pair of different nodes, is called a tournament. The number of possible tournaments increases exponentially with the number of players, as each arrows orientation changes the graph, resulting in a separate tournament. For n nodes there are exactly $\binom{n}{2} = \frac{n(n-1)}{2}$ many arrows, as there are that many 2-element subsets of the set $\{1, \dots, n\}$, meaning that this many links are necessary to connect each pair of different nodes. Because each of the connecting arrows can have two orientations, there are $2^{\binom{n}{2}}$ tournaments of size n (that is, n players). This sequence grows rapidly: for $n = 2,3,4,5,6,7$ players there are $2, 8, 64, 1024, 32768, 2097152$ many tournaments respectively. More elements of this sequence can be retrieved from the *On-Line Encyclopedia of Integer Sequences* under sequence number A006125 *(Sloane, 2018)*.

(Note that this is also the number of undirected graphs on n nodes; however, a number of these graphs will be disconnected.) Below, *figure 15* displays one of the 32768 possible tournaments for six nodes.

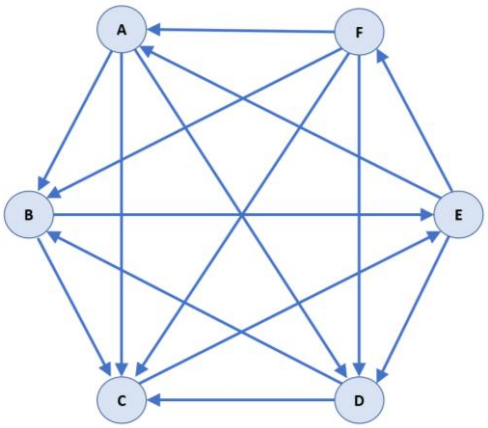


Figure 15. Network depicting one possible outcome of a graph-theoretical tournament between six nodes. Each pair of nodes is connected, the arrow pointing towards the winner of each match.

In real-world applications, such tournaments occur in *round-robin* settings, a type of tournament, where each pair of players competes in an individual match in order to determine an overall winner.

Curiously, groundwork was laid in this field, when H. G. Landau (1953) studied dominance relations and the societal structure of chickens in a flock. His real-life observations allowed him to mathematically define key properties of tournaments like the number of wins in a tournament being equal to the number of games and that number being $\frac{n(n-1)}{2}$ when he proves his theorem on a score structure *(Landau, 1953)*.

It is important to keep in mind that abstract mathematics can be inspired by real-life circumstances but that in turn, abstract mathematics may have unexpected applications in the "real" world – with all corresponding implications that can arise from social-networks.

4. Measures of centrality

There are numerous measures of centrality, a selection of which are degree centrality, closeness centrality, eigenvector centrality, betweenness centrality, the rush index, flow betweenness, information centrality and more (Borgatti, 2005; Bonacich, 1987; Freeman, 1979; Freeman, 1977). Above, the general notion of centrality was introduced as a concept of more or less advantageous locations within a network. Below, four measures of centrality are described in detail: degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. Three of these are discussed by Freeman (1979), who in fact discovered betweenness centrality, all four are identified as important by Borgatti (2005) and commonly used in the analysis of networks (Borgatti, 2005). After this introduction we will further study these four concepts by means of selected research papers. Borgatti (2005) stresses the importance of using the right

measure, compatible with the characteristics of the studied network. He warns: "*It is shown that the off-the-shelf formulas for centrality measures are fully applicable only for the specific flow processes they are designed for, and that when they are applied to other flow processes they get the 'wrong' answer*" (Borgatti, 2005, p. 55)

All measures of centrality discussed in this work are measures of *point centrality* (as opposed to *graph centrality*). This means that we are measuring characteristics of points and not of entire graphs (Freeman, 1977).

Let us now define these measures.

4.1 Degree Centrality

The basic idea of degree centrality is simply that of counting the number of edges leading to a node (Borgatti, 2005). A number of special kinds of degree centrality have been developed that are applicable only under specific circumstances (such as special network structure or adjacency requirements). For examples of these, see Rogers (1974) or Czepiel (1974). Nieminen (1974) however formalized the general measure of degree centrality for undirected graphs as the number of adjacent nodes i to a given node j:

$$C_D(j) = \sum_i a_{ij}. \tag{1}$$

a_{ij} can either be 1, if the nodes i and j are connected or 0, if they are not (Nieminen, 1974). $C_D(j)$ is 0, if j is isolated and not connected to any other nodes, and it is large, if j is connected to many nodes. We will come across this measure several times, as it is such an obvious and natural object.

In some cases, it may be useful to normalize the quantity $C_D(j)$ to be able to compare similar networks of differing size:

$$C'_D(j) = \frac{C_D(j)}{n-1}, \tag{2}$$

where n is the number of nodes in the network (Freeman, 1979). Division by $n-1$ is sensible because it describes the maximum number of edges possible, the maximum $C'_D = 1$ is attained for the star graph, for example (see *figure 16*).

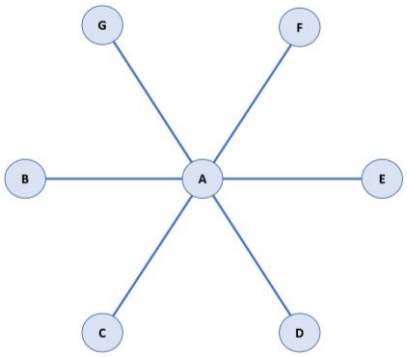


Figure 16. Star graph with n=7 nodes and n-1=6 edges. Since a_{ij} is either 1 *or* 0 *and* $0 \leq C_D(j) \leq n$*, division by* $n - 1$ *can leave* C_D' *no larger than* 1.

This normalization is helpful, if the relation between the number of nodes and edges is comparable among examined networks. Also, the normalization leads to C_D' converging towards 0, if the average number of connections is small compared to network size (consider Facebook, where each users' number of "friends" is miniscule compared to the total number of clients).

Borgatti (2005) calls degree centrality a "*measure of immediate influence*" (Borgatti, 2005, p. 62) because only adjacent nodes affect each other, and distant edges have no influence in the measure of a node. In terms of flow, a possible realization is parallel duplication. Consider a randomly distributed and copiable good that travels with a certain probability from one node to the next in a network. Each node's probability of receiving this good is dependent on the number of his/her immediate neighbors as each node has an equal probability of having the good in the first place (Borgatti, 2005). Another application for degree centrality given by Borgatti (2005) is the number of times a node is visited by a good taking a random walk (both nodes and edges can be visited multiple times) through the network. The greater the number of arrows pointed toward a node, the greater the probability of a random walk leading to (or through) that node – as discussed above, these properties are measured by degree centrality.

Degree centrality can be calculated very easily and is therefore suitable for everyday use. Possible applications in an economic context include tracking in-and out-going account activity for book keeping purposes or monitoring transactional records. Among wider applications are the identification of individuals who are likely to have direct access to a big number of people in order to quickly spread ideas, advertisement or other goods that need to be distributed quickly. Also, individuals with a high degree centrality might be interesting first-contact targets for example for companies trying to establish new markets because they are likely able to gather community feedback quickly – and efficiently.

4.2 Betweenness centrality

The second measure of centrality we wish to consider is betweenness-centrality. The rough idea of this measure is that a node's importance increases with the number of geodesics (shortest paths) it lies on. Freeman (1977) introduced this measure of centrality as "*…the degree to which a point falls in the shortest path between others and therefore has a potential for control of communication.* [It] *… may be used to index centrality in any large or small network of symmetrical realltions, whether connected or unconnected*", (Freeman, 1977, p. 35). An important premise is that goods are indivisible and thus can only travel along one path at a time. Moreover, every possible (shortest) path is selected with equal probability (Freeman, 1977); (Borgatti, 2005).

Consider all shortest paths from A to F in the graph depicted in *figure 17*. There are three such paths: $A-D-G-F$, $A-C-E-F$ and $A-D-E-F$. Node D lies on two of these three geodesics, meaning that the pair (A,F) contributes $\frac{2}{3}$ to the betweenness centrality of D. (Node G on the other hand lies on only one of these three paths, thus receiving $\frac{1}{3}$ as (A,F)'s contribution; B on the other hand, which is on no shortest path, receives 0.). To calculate D's total betweenness centrality, we have to sum up all contributions of pairs (X,Y), where $X \neq Y$ and $X \neq D \neq Y$. In plain language, we sum up all contributions of all shortest paths D lies on.

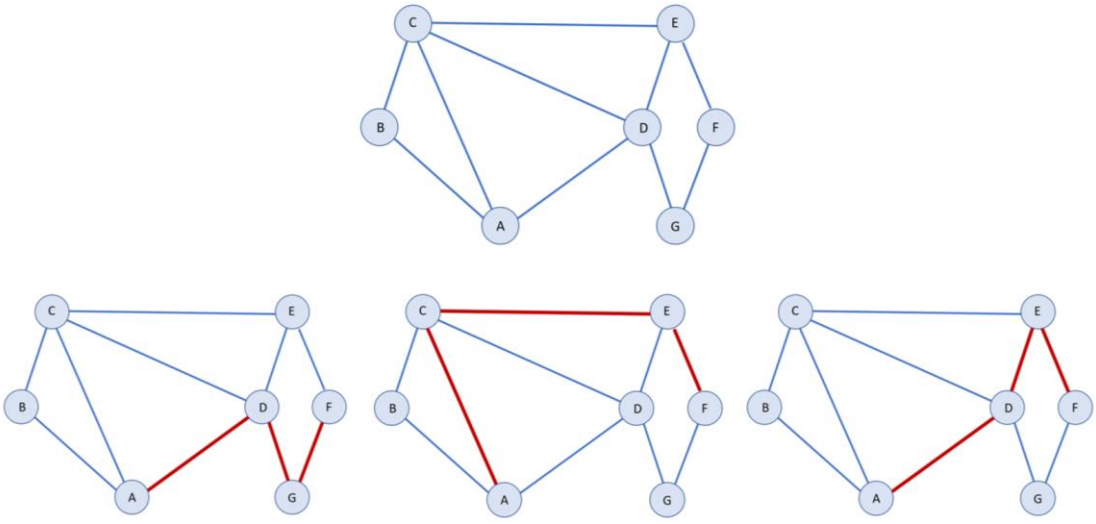


Figure 17. Depiction of all three shortest paths (geodesics) between nodes A and F.

Formally, as described by Borgatti (2005), who uses a slight variation of Freeman's (1977) notation, betweenness centrality is defines as follows. Let g_{ij} be the number of geodesics between nodes i and j. Assuming that all shortest paths are chosen with equal probablility and

that only one such path is selected at any time, the probability for coosing a particular path is $\frac{1}{g_{ij}}$ under the assumption that $g_{ij} \neq 0$. By g_{ikj} he denotes the number geodesics crossing k. Then the probability that node k lies on a randomly chosen shortest path between i and j is $\frac{g_{ikj}}{g_{ij}}$. Adding up these probabilities for all i,j such that $i \neq k \neq j \neq i$, we receive the total betweenness centrality $C_B(k)$ of a node k:

$$C_B(k) = \sum_{i} \sum_{j:\, i \neq k \neq j \neq i} \frac{g_{ikj}}{g_{ij}}. \tag{3}$$

Note that g_{ij} may be 0 if there is no connection between i and j. In such a case g_{ikj} logically needs to be 0 as well. In the above formula (3), we assume that $\frac{0}{0} = 0$. This property is implicitly assumed, but not stated explicitly in neither Freeman (1977) nor Borgatti (2005). Radicchi (2011), who uses the same assumption similarly, also does not deal with this case explicitly.

Below, the betweenness centrality for the graph depicted in *figure 17* is calculated. Rows correspond to k, the node we want to calculate the centrality measure for, and columns indicate pairs i,j of nodes between which the shortest paths are considered. Observe that D has the greatest betweenness centrality of this graph, while B has the lowest. The ranking induced is $D > C > E > A > G > F > B$.

Table 1. Betweenness centrality of nodes A to G from the network depicted in figure 17.

	AB	AC	AD	AE	AF	AG	BC	BD	BE	BF	BG	CD	CE	CF	CG	DE	DF	DG	EF	EG	FG	$C_B(k)$
A	0	0	0	0	0	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0	**1**
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
C	0	0	0	$\frac{1}{2}$	$\frac{1}{3}$	0	0	$\frac{1}{2}$	1	1	$\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0	$^{23}/_{6}$
D	0	0	0	$\frac{1}{2}$	$\frac{2}{3}$	1	0	0	0	0	1	0	0	0	1	0	0	0	0	$\frac{1}{2}$	0	$^{14}/_{3}$
E	0	0	0	0	$\frac{2}{3}$	0	0	0	0	1	0	0	0	1	0	0	$\frac{1}{2}$	0	0	0	0	$^{19}/_{6}$
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{2}$	0	$^{1}/_{2}$
G	0	0	0	0	$\frac{1}{3}$	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{2}$	0	0	0	0	$^{5}/_{6}$

Betweenness centrality is all about shortest paths and the flow along them. Economic applications therefore include situations where quick flow is required. Imagine a fashion blogger with high betweenness centrality, who might be able to direct a trend (and sales) towards a certain brand or product when respective sales go down. This would be possible because he could

influence the speed with which a trend travels. For the same reason, betweenness centrality comes with the power for disruptive behaviour and a node with a high betweenness measure has the ability to cause flow to stagnate or possibly even to fail in time-sensitive situations. Because of these characteristics, nodes with a high betweenness centrality often act as brokers or are consulted as such by other members of their network, be it to broker business dealings, information, work force or any other good that can be exchanged along the network at hand.

4.3 Closeness Centrality

The third measure of centrality we wish to examine is *closeness centrality*. When Bavelas (1950) introduced his precursor to modern closeness centrality, he developed a method that lends itself beautifully to induce a ranking. Bavelas' objective was to study communication patterns: their efficiency and effectivity in task-oriented groups (Bavelas, 1950), He found that in real-world situations social processes within groups drive communication paths (even if different paths are formally deemed superior). However, he states that "*...any hope of success depends upon an effective flow of information*" (Bavelas, 1950, p. 725). The study of patterns of graphs and distance between nodes led to his definition of *dispersion* and *relative centrality*. "Dispersion" signifies the "sum of internal distances" $\sum d_{ji}$ between nodes i and j. For the network depicted in *figure 18* below (based on Bavelas, *figure 2*, p. 276), the dispersion figure is 20 $(6 + 4 + 4 + 6)$.

Figure 18. Simple linear network with four nodes A-B-C-D.

Table 2. This table displays the sum of distances $d_j = \sum_{i:\, i \neq j} d_{ji}$ of points A, B, C and D to all other points.

AB	1	BA	1	CA	2	DA	3
AC	2	BC	1	CB	1	DB	2
AD	3	BD	2	CD	1	DC	1
d_j	6		4		4		6

Under "relative centrality", we understand the "ratio of the sum of all internal distances to sum of distances from a particular position" (Bavelas, 1950):

$$\frac{\sum d_{ji}}{d_j} \tag{4}$$

For the nodes B and C in the above network, this measure is $\frac{20}{4} = 5$, whereas for nodes A and D it is the smaller value $\frac{20}{6} = 3.\dot{3}$.

Beauchamp (1965) refined the notion that information could be spread very successfully through a network, if nodes have short distances to other nodes. He defines point centrality of a node j as follows:

$$C_C(j) = \frac{n-1}{\sum_{\mathrm{i:\, i \neq j}} \mathrm{d_{ji}}}. \tag{5}$$

We wish to generalize our four-point example from above to a line of n nodes.

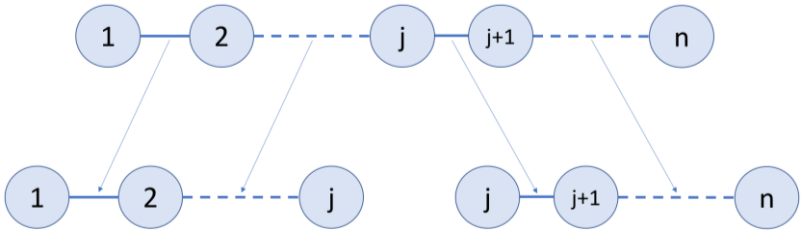


Figure 19. We split the line graph at node j to simplify the problem of summing up all d_j.

In order to calculate $C_C(j)$, we face the problem of calculating d_j, that is the problem of summing up all distances of j to all other nodes $i \neq j$. We wish to employ the sum formula $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$ but need to make adjustments, because we need the formula to take care of general positions j that are variable. To simplify the problem, we split the graph at the the j^{th} node in order to obtain two graphs, one comprising the nodes 1 up to j, the other containing the nodes j to n. This way, we have reduced to problem to two simpler problems. The first is computing $\mathrm{d_j}$ of the first graph, which is the same as computing $\mathrm{d_1}$ of the same graph, because of the symmetry of the problem. The second is computing $\mathrm{d_j}$ of the second graph (nodes j to n). The second is more complex, because we need to rename the nodes by shifting the indices by $j-1$. The second problem is therefore the same as computing $\mathrm{d_1}$ for the graph with nodes 1 up to $n-j+1$ (obviously, we are considering a different graph). We therefore need to compute $\mathrm{d_1}$ for the graphs with nodes $1, \dots, L$, where $L = j$ (for the first graph) and $L = n-j+1$ (for the second graph). Now we can use the sum formula from above: for such a graph, the sum of distances from node 1 is $1 + 2 + \cdots + (L-1) = \sum_{i=1}^{L-1} i = \frac{L(L-1)}{2}$. Consequently, using this formula twice, multiplying everything out and regrouping terms, we receive

$$\mathrm{d_j} = \frac{j(j-1)}{2} + \frac{(n-j+1)(n-j)}{2} = \frac{n(n+1)}{2} - j(n-j+1). \tag{6}$$

We have used the above formula (6) to calculate $C_C(j)$ for a linear graph with n =100 nodes. In *figure 20* we see that central nodes have the highest closeness centrality (as is expected) and that the distribution of values is symmetrical.

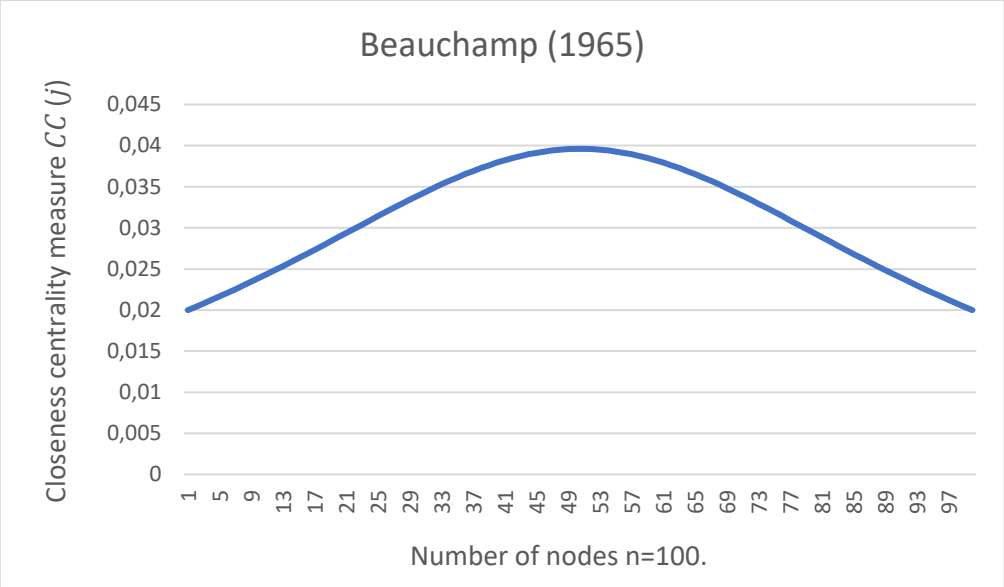


Figure 20. Symmetrical distribution of closeness centrality $C_C(j)$ in a line graph with n=100 nodes.

We wish to compare the above formula (6) for d_j to a formula for betweenness centrality of points in a graph as above (linear graph with n nodes). Let us compute $C_B(j)$ as defined in Borgatti (2005). Suppose that $1 \leq j \leq n$ (see *figure 21*) with nodes i, l such that $i \neq j \neq l \neq i$.

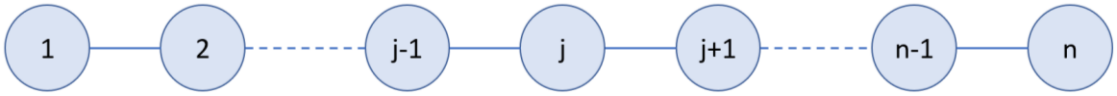


Figure 21. Line graph with n nodes.

To calculate $C_B(j)$, we need to sum up all geodesics running through j. Logically this applies only to shortest paths between nodes i and l, if they lie on opposite sides of j. Moreover, because we are considering an undirected graph both shortest paths $i \to l$ and $l \to i$ exist. We therefore assume that i is to the left side of j with l lying to its right and multiply the calculated shortest paths by two. There are $n - j$ shortest paths $i \to l$ running through j for every node $i < j$. (There are also $j - 1$ shortest paths $l \to i$ running through j for every node $l > j$.) We can therefore multiply the number of nodes on the left of j by the number of nodes on its right and obtain

$$C_B(j) = 2(j-1)(n-j). \qquad (7)$$

This expression resembles the second term in our formula (6) for d_j (the other term does not depend on j). For completeness, we also give a plot of the betweenness centrality for the linear graph with $n = 100$ nodes.

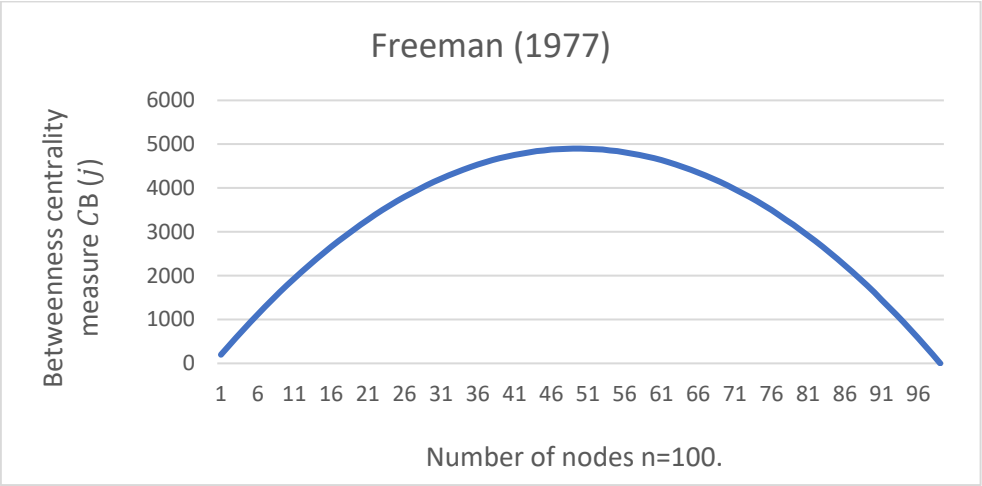


Figure 22. Symmetrical distribution of betweenness centrality $C_B(j)$ in a line graph with n=100 nodes.

Note that these values are twice the size as values obtained by the original Freeman (1977) formula, obviously the shape of the graph is unchanged. In this very special case, the rankings obtained by the different measures of centrality are the same, as is obvious when looking at the plots: highest ranks are achieved by the middle nodes, lowest ranks by the outermost nodes. That is, 50 and 51 are in the first rank, 49 and 52 are in the second, and so forth.

The application of this measure of centrality centres on speed and range. An individual with a high measure of closeness centrality has the ability to quickly spread information (or other goods) throughout the entire network. If, say, a position needs to be filled very quickly, an actor with a high betweenness measure will likely be able to broadcast this information well. Also, because of the ensuing reach, individuals with high betweenness centrality may be powerful influencers, simply because they have better access to the network as a whole most quickly. Another important feature is that their power can be interpreted in terms of independence. An individual who has broad reach will be less dependent on brokerage through other actors.

4.4 Eigenvector centrality

Finally, we wish to illustrate the fourth measure of centrality: eigenvector centrality. As opposed to all other notions of centrality introduced so far, this measure does not allow the evaluation of individual nodes. Instead, the relevance of each node is dependent on the relevance

of the neighbouring nodes – which are also unknown at the onset. We are therefore dealing with a system of equations where each node corresponds to one variable (Wasserman & Faust, 1994). To formalize this idea, Bonacich (1972) used the following approach: We are considering a directed graph, where $j \to i$ means that the tie between i and j is directed towards i. Let P_i denote the valuation of node i. This valuation P_i should depend on the valuation P_j of neighbouring nodes j. More precisely:

$$\lambda P_i = \sum_{j:j \to i} P_j, \tag{8}$$

where λ is a parameter to be chosen (a so-called *eigenvalue*). (This P_i is the centrality we wish to calculate to achieve an assessment of each node's importance.) Note that $P_i = 0$ for all i is always is a solution, but we are not interested in this solution. Instead, we wish to find the greatest possible λ.

In order to obtain a familiar notation that is used in mathematics and physics, we define a matrix W with the entries w_{ij}. If $j \to i$ we set $w_{ij} = 1$, $w_{ij} = 0$ otherwise. The resulting matrix is an adjacency matrix for the graph we are considering. Using this notation, we can now rewrite equation (8) as follows:

$$\lambda P_i = \sum_j w_{ij} P_j \tag{9}$$

in other words

$$\lambda P = WP \tag{10}$$

where P is the vector consisting of the entries P_j. In a more extensive form, this can be written as

$$\lambda \begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix} = \begin{pmatrix} * & \cdots & * \\ w_{i1} & \cdots & w_{in} \\ * & \cdots & * \end{pmatrix} \begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix}. \tag{11}$$

Equation (10) is an *eigenvalue equation*, value λ *is an eigenvalue* and the vector P is an *eigenvector* (if it is not 0) (Franks, Griffith, & Anand, 2014; Bonacich P. , 1972; Wasserman & Faust, 1994). Further it is known that only the largest eigenvalue gives sensible results because for all other λ there is at least one entry P_i in the eigenvector that is not a real number ≥ 0 (Frobenius, 1912).

Summarizing these observations, we can state that the vector of eigenvector centralities P_i is an eigenvector of the adjacency matrix.

Returning to plain language, eigenvector centrality can only be solved for all nodes of a network at once. This is due to all values being mutually interdependent. An actor achieves a high

centrality score if many actors with high scores have arrows pointing towards him. In other words, each node's importance is dependent on it's neoghbours importance (Borgatti, 2005). Ranges for the application of eigenvector centrality and its derivatives are wide as we believe that it can often be used as a "fallback" measure of centrality. For example it would be interesting to study the usefulness of this measure in networks derived from round-robin tournaments as this could help dealing with the problem of ties. Most famously, the PageRank-algorithm, which the web-search engine Google is based on, is a variant of eigenvector centrality (Brin & Page, 1998). Eigenvector centrality has also been used to predict the social rise of male birds (lek-mating wire-tailed manakins, to be precise) (Ryder, McDonald, Blake, Parker, & Loiselle, 2008). However, there may be more economic value in applying this measure, for example, on the study of relationships in social media networks.

5. Discussion of literature and simulation

In the following section four scientific articles are discussed to illustrate each measure of centrality's application in contemporary science. Chapter 5.4 contains a self-devised implementation of a diffusion-based algorithm.

5.1 Degree Centrality: Ranking the participants in a tournament – (Rubinstein, 1980)

An elementary method of ranking the participants in a tournament network, called the "points system" was studied by Rubinstein (1980). Given a network derived from a complete tournament, a ranking of participants can be obtained by counting each participants' number of wins. This is the basic idea of the points system ranking method.

There is a number of prerequisites that need to be considered:

1. We are interested not in ranking any arbitrary network, but specifically in ranking participants of networks resulting from tournaments.
2. Tournaments consist of $\binom{n}{2} = \frac{n(n-1)}{2}$ matches (each pair of players competing), each match resulting in a unique outcome of one winner and one loser which are represented by arrows.
3. The result of any tournament is given – meaning the network itself is given.

In his paper Rubinstein (1980) defines a set of three axioms that completely characterizes this ranking method: the points system satisfies all three axioms and is the only system to do so.

This formal approach to a seemingly simple method has two distinct merits: Axiom I is a desirable characteristic for ranking methods of different complexity, Axioms II and III highlight the fact that rankings are based on local network properties.

The three axioms, as defined by Rubinstein (1980), are:

Axiom I: "*Anonymity Axiom*": renaming any participant does not influence the ranking.

Axiom II: Suppose i and j are distinct players in a tournament and are ranked such that $i \geq j$. Assume that a third player k wins against i, that is, $k \to i$. If this arrow is reversed, yielding a new tournament where $i \to k$, then i is ranked strictly greater than j, in symbols $i > j$.

Axiom III: Suppose that i, j, k and l are distinct players. The outcome of the match between k and l does not influence the relative ranking of i and j.

The following example illustrates the points system as well as the above axioms:

Consider the tournament (*figure 23*) given by the six $(= 4(4-1)/2)$ beating relations $a \to b, a \to d, b \to c, b \to d, c \to a, d \to c$. Players a and b both win two matches, while c and d only achieve one win respectively. The points system allocates a ranking corresponding to these numbers, resulting in a and b coming in first place/ rank, and c and d sharing second place/ rank. The fact that a beats b does not prevent both from sharing the same rank. The fact the c beats a does not cause c to be placed higher or as high as a, in fact c is ranked below a.

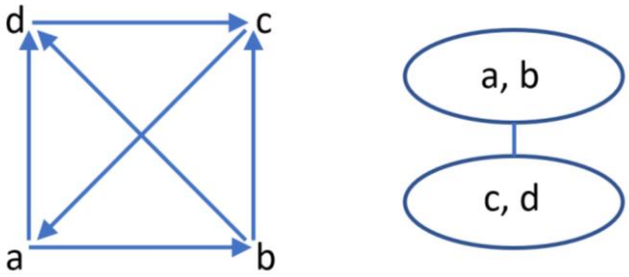


Figure 23. Tournament between nodes a, b, c and d. Arrows indicate winners of individual matches. To the right the resulting ranking achieved through the points system.

We wish to understand the axioms by means of this example. Axiom I states that renaming any participant of the tournament does not influence the players rankings. If we were to change the players designations in Figure 1 from a, b, c and d to w, x, y and z, the ranking induced by the ranking method would be obtained by applying the same substitution of monikers to the original ranking. Instead of a and b we would find w and x on top rank and y and z on the bottom rank. In order to understand Axiom II, consider the three players a, b, c. Players a and b occupy the same rank, but as soon as we let a win against c, all else unchanged, player a occupies a higher rank than player b.

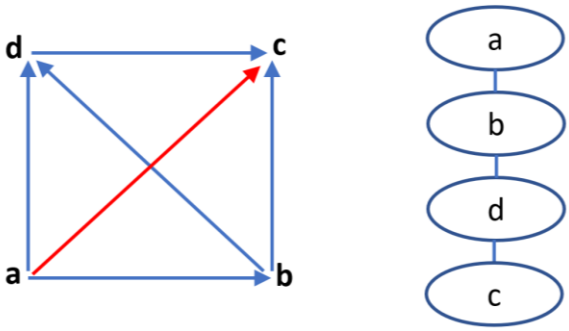


Figure 24. Tournament as seen in figure 23 with the difference that the match between a and c is won by a. To the right the resulting ranking achieved through the points system.

Finally, changing the outcome of the match between c and d does not change the ordering of a and b in the ranking, even though they now share the top rank with c.

All three axioms are satisfied only by the points system, rendering it one of the most simple ranking methods.

Most networks found in the "real world" are not complete (containing pairs of nodes not directly connected) and rankings for naturally occurring networks are the subject of this work. It is therefore

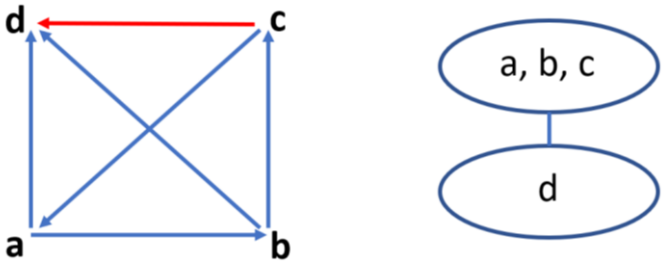


Figure 25. Tournament as seen in figure 23 with the difference that the match between c and d is won by. To the right the resulting ranking achieved through the points system.

an obvious question if the points system can be extended to general networks: Is the points system applicable to non-complete networks and does this ranking method satisfy Rubinstein's axioms I, II and III for such networks? The former is obviously the case, yes. Counting the number of outgoing arrows (wins) is just as possible for non-complete networks as it is for complete networks. The latter can be verified as easily as in the complete case: Axiom I is valid because names do not influence the number of outgoing edges. In order to verify Axiom II, we consider two nodes i and j such that i has at least as many outgoing arrows (wins) as j. Replacing the arrow $k \to i$ by $i \to k$, the number of wins of i is strictly larger than that of j. In the same way, Axiom III can be verified: the orientation of the arrow between k and l does not

influence the number of outgoing arrows of i and j. Moreover, in applications, draws may occur, necessitating the possibility of weighted outcomes. We therefore also consider directed graphs with weighted edges, to allow for such possibilities. Clearly, instead of the number of wins, we are now interested in the sum of weights of outgoing edges. Repeating the arguments from above, we see that Axioms I, II and III again hold true (for weights > 0).

Potential applications

Because the points system is such a simple and easy to apply method, it lends itself to common use. Firstly, round-robin tournaments are an ideal premise for the application of the points system, as the structure of such tournaments renders complete networks. For example, preliminary rounds of the football World Cup and European Championship use the round robin format (a knock-out system is used for the main round) (Müller, 2017).

However, large numbers of participants imply an increasing number of matches ($n(n-1)/2$), causing ever greater effort and investment for tournament organizers and participants. Where 5 players require 10 matches, 10 participants already necessitate 45 matches and so on. This implies that round robin tournaments may only be of use for limited participant numbers or where a great number of matches does not correspond with greatly increased cost.

Secondly, the points system may be applied to more general networks, but it is unclear if this is sensible in all cases. Consider the non-complete network a, b, c, d, e, f, g, h with loosely connected areas of high density A and B: area A consisting of players a, b, c and d, area B consisting of players e, f, g and h. If all players in area B are better than the players in area A, such that any player in B would beat any player in A, the points system will not lead to a meaningful ranking. Players within the weaker area A are able to achieve an equal or higher number of wins against each other, than players in the stronger area among each other. This leads to weaker players achieving an equal or higher rank than players of the stronger area B. *Figure 26* depicts a possible scenario:

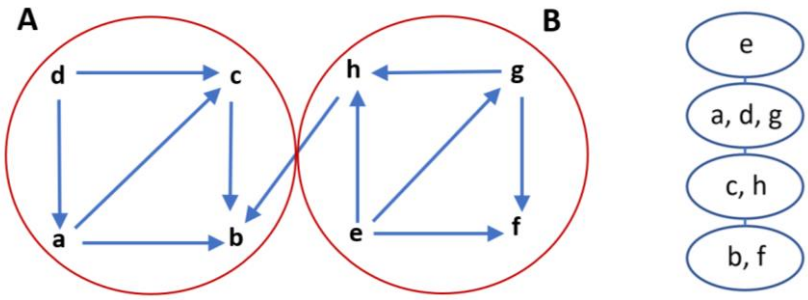


Figure 26. Non-complete network with two separate areas of high density. If all players in area B are better than the players in area A, the points system does not lead to a meaningful ranking.

This demonstrates that the points system, while easily applicable, is not suitable in all cases. Also, due to the method of counting outgoing (weighted) arrows, there may not always be a distinct winner, but several participants might share each rank.
Below, a diffusion algorithm developed by Radicchi (2011) will be introduced. For knock-out tournaments, that method yields the same ranking as the Points System.

5.2 Betweenness Centrality: Quantifying the Performance of Individual Players in a Team Activity – (Duch, Waitzman, & Amaral, 2010)

Many real-world networks consist not of independent individuals, but of teams working together to achieve a common goal. This is true not only in sports but in business settings, science and social communities as well. Big companies such as Google or Facebook actively add teamwork skills and opportunities into their job offers, making the ability to work well in teams a crucial recruitment requirement. Targeted team-selection and -composition can be crucial to establish maximum output or increased creativity. Within a team, certain roles will be occupied that are often not interchangeable and whose impact cannot be easily compared. In football for example, there are goalkeepers, centre forwards and defence players, all are necessary to successfully compete in a match, but their contributions are of different nature. Measuring the performance of a team is obviously dependent on its members' contributions but these individual contributions need to be measured with the different assignments in mind, necessitating individual performance evaluations. This allows for the interesting question if (large) discrepancies of compensation of players within a team are warranted and if such differences in pay arise from actual differences in those player's contribution (Duch, Waitzman, & Amaral, 2010). Diversity and a lack of redundancy are looked-for to achieve a balanced and fruitful collaboration basis in a network. In terms of network-structure, it is desirable to obtain a composition that allows for easy information (or other desired commodity) access along shortest possible paths.
Duch, Waitzman, & Amaral (2010) study performance evaluation of teams and individuals on the basis of the 2008 Euro Cup football tournament. By utilizing large amounts of publicly available statistical data, they devise a variant of betweenness-centrality, called "flow centrality", that *"…captures the fraction of times that a player intervenes in those paths that result in a shot"* (Duch, Waitzman, & Amaral, p. 2, 2010). While betweenness-centrality is concerned only with shortest paths between two nodes, *flow centrality* as defined by Duch, Waitzman, & Amaral (2010) also takes into account measures of passing and shooting accuracy of each player, thus extending the original concept. The resulting diagrams are complex visualizations

of each player's contribution in a given match, depicting both passing accuracy and individual performance as well as the number of successful passes.

Figure 27 shows how Duch, Waitzman, & Amaral (2010) depict the studied matches in terms of flow networks. Node size is contingent on player performance and colour coded by passing accuracy; arrow thickness represents the number of successful passes between two players, arrow colour codes arc centrality.

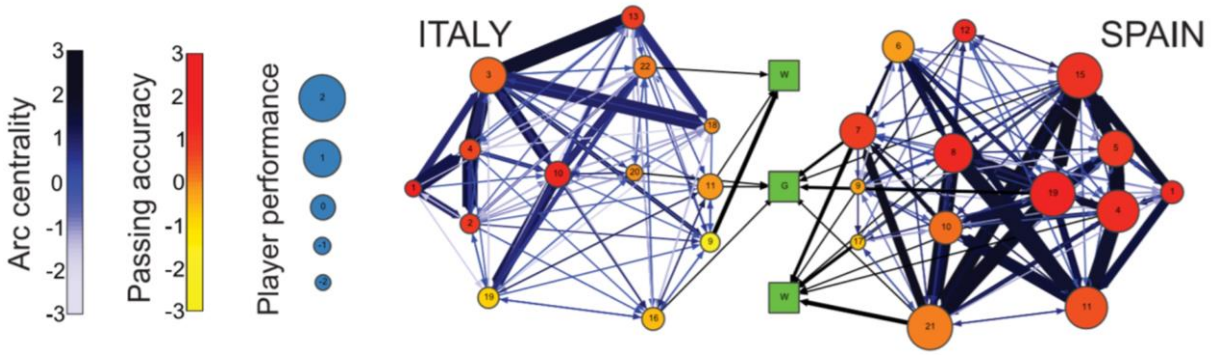


Figure 27: Rearranged from Figure 5 in Duch, Waitzman, & Amaral 2010. The flow network depicts centrality, passing accuracy and player performance of studied matches.

Let N denotes the total number of all players in a team. Each player i of team A is assigned a match performance ρ_i^A according to his flow centrality measure. For $n \leq N$ team performance $\rho^{A,n}$ is measured as the average performance of the best n individual players (sorted from best to worst) of a given team A:

$$\rho^{A,n} = \frac{1}{n} \sum_{i=1}^{n} \rho_i^A. \tag{12}$$

These values offer an opportunity to compare the performance of any pair of teams by calculating the difference δ_n of team performances:

$$\delta_n = \rho^{A,n} - \rho^{B,n}. \tag{13}$$

The value of this measure lies in its predictive power. Duch, Waitzman, & Amaral (2010) find significant correlation between the values δ_2 and the actual outcome of a match. In fact, they suggest a significant correlation for values of $n \leq 4$ (that is a correlation for the case of four or fewer than four of the best players of a given team). This finding is of great interest, as it implies that only the top players' performances are determining factors for match outcomes. This in turn can be viewed as justification for compensation disparities within teams. If it is only the top performing players who are the deciding factor to lead their team to victory, compensation premiums seem justified.

This method of ranking individuals in a network, though applicable in sport, was not proven viable in a tentative study in the context of teamwork within scientific projects (Duch, Waitzman, & Amaral, 2010). However, further and more detailed research is warranted to expand on possible additional applications of this approach like the above mentioned identification and ranking of individuals with the power to exert influence within their network.

5.3 Closeness Centrality: Fast Approximation of Centrality – (Eppstein & Wang, 2004)

A common problem when calculating centralities is that due to increasing network size, computations get ever more complex. While this is not a problem in principle, even for most personnel computers, vast data ranges require a long time to process. With time being a valuable commodity, expediting the computing process is welcomed.

In computer science, the quality of an algorithm is given by the time it takes to solve a problem. This can be measured by the number of elementary computations, as this is what computers are basically capable of. The notation O is used to measure the complexity of an algorithm. For example, if the complexity of an algorithm is $O(n^2)$, this means that if the input size is n (for example number of nodes in a graph) the algorithm needs at most Cn^2 operations to solve the problem where C is constant. Consider the multiplication of two numbers using pen and paper. For single digits the multiplication requires one step. Multiplication of larger numbers, say $87 \cdot 23$, involves multiplying each digit of the first number with each digit of the second number. In the above case four elementary multiplications are needed. For numbers having n respectively m digits, the number of elementary multiplications is $n \cdot m$. This usual "pen and paper algorithm" therefore has complexity $O(nm)$.

There are different algorithms used at present to compute closeness centrality in undirected graphs (Eppstein & Wang, 2004). One possibility to solve this problem is through an algorithm that takes into account all nodes n and delivers a result for all of these at once at the end of the computing process. There are no interim results in such a case and to obtain any particular closeness centrality (for a selected j), the computation has to be finished. One such algorithm solves $C_C(j) = \frac{n-1}{\sum_{i:\, i \neq j} d_{ji}}$, which is equation (5) by Beauchamp (1965) above, for all j in time $O(nm + n^2 \log n)$ where m is the number of edges and n is the number of nodes (Johnson, 1977).

The other general path to solve equation (5) is to use an algorithm that calculates $C_C(j)$ for individual j (one at a time). This obviously can be used to calculate the closeness centrality for

all nodes j but does not need to. Solving the *Single Source Shortest Path* (SSSP) problem is one way to do this. An algorithm that can do so in time $O(n \log n + m)$ was discovered in by Fredman and Tarjan in 1987 (Fredman & Tarjan, 1987). To recap, this means that it takes time $O(n \log n + m)$ to compute the closeness centrality $C_C(j)$ for any individual node in a network consisting of n nodes and m edges. To compute $C_C(j)$ for all j using this algorithm, it takes n times as long and curiously we arrive back at Johnson's estimate $O(nm + n^2 \log n)$ (Johnson, 1977).

Eppstein and Wang (2004) have also tackled this latter problem. Their approach is to give an approximation of $C_C(j)$ for any node by using a randomized algorithm. First, they select k nodes $v_1, \ldots, v_k$ at random (k to be selected according to the size of a network and desired quality of approximation) and calculate their respective distances to all nodes, using the algorithm developed by Fredman &Tarjan. *Table 3* shows an example for such a table of distances, with rows corresponding to the pivot points and columns indicating all nodes. The time needed to compute this table is $O(kn \log n + km)$. Adding up all entries in a row would yield the closeness centrality $C_C(v_i)$, but this is not what we want.

Table 3. Table of distances between the pivot points v_i and all points of the network. In this abstract example $k = 3$ and $n = 6$.

	A	**B**	**C**	**D**	**E**	**F**
$\boldsymbol{v_1}$	$d(A, v_1)$	$d(B, v_1)$	$d(C, v_1)$	$d(D, v_1)$	$d(E, v_1)$	$d(F, v_1)$
$\boldsymbol{v_2}$	$d(A, v_2)$	$d(B, v_2)$	$d(C, v_2)$	$d(D, v_2)$	$d(E, v_2)$	$d(F, v_2)$
$\boldsymbol{v_3}$	$d(A, v_3)$	$d(B, v_3)$	$d(C, v_3)$	$d(D, v_3)$	$d(E, v_3)$	$d(F, v_3)$

Instead, they consider the columns of this table. By adding up the values in column j and multiplying the obtained value by $\frac{n}{k}$, they obtain an estimate for the sum of distances from j to all other nodes. This can be done because the selection of random k nodes usually is a good representative of the entire network. The result is further used to obtain a centrality estimator for node j as follows:

$$\hat{C}_C(j) = \frac{n-1}{\frac{n}{k}\sum_{i=1}^{k} d(j, v_i)} \tag{14}$$

Eppstein & Wang (2004) could reduce the time needed to calculate (by approximation) the closeness centrality for all nodes j to $O(\frac{\log n}{\varepsilon^2}(n \log n + m))$ with ε being the quality of approximation (the smaller ε, the better the approximation). However, there are two caveats: firstly,

$\hat{C}_C(j)$ is only an approximation of $C_C(j)$ and secondly, the algorithm yields the desired result only with the high probability of $1 - \frac{1}{n}$ because of the random selection of the pivot nodes.

Let us now roughly compare the complexity of Johnson's algorithm with that of Eppstein and Wang. We select a network with 10^9 nodes and m edges. Johnson's method yields a running time of $O(nm + n^2 \log n) = O(n(m + n \log n))$. Eppstein and Wang's approximation returns a result in $O(\frac{\log n}{\varepsilon^2}(n \log n + m))$. Note that the complexities differ only in the first factor, it is therefore sufficient to compare these facors: n versus $\frac{\log n}{\varepsilon^2}$. For $\varepsilon = \frac{1}{100}$ the quotient of these two quantities is about 5000. To put this into perspective, a calculation that would take one year (8760 hours) with the first algorithm is finished within a couple of hours when using the approximation method.

5.4 Eigenvector Centrality: Who is the best player ever? A complex network analysis of the history of professional tennis. – (Radicchi, 2011)

In some cases, tournament networks demand sophisticated ranking methods to do justice to the fact that each node's ranking should depend on the ranking of its neighbours. Winning against a strong participant should be valued more highly than winning against a weak opponent. This property allows a more nuanced ranking and is the main feature distinguishing such a system from the more basic points system.

Radicchi (2011) developed a diffusion algorithm that allows to assign a score to each node, based on the value of all other nodes in the network. In other words, the scores are determined by solving a system of equations in N variables, where N is the number of nodes. This algorithm is heavily built on Google's "PageRank" system, developed by Brin and Page (1998).

The general idea is to determine the best tennis player in history, by analysing all results from certain professional tennis matches (Grand Slams and ATP World Tour tournaments) taking place between 1968 and 2010. In order to do so, a complex network of tournaments needs to be established that has the capacity to compare players of several generations. He also needs to cover the fact that not all pairs of players have competed and that a great number of actual tournaments needs to be considered. Interestingly, the only necessary input is information about which pairs of players competed and who won each match, no external evaluations (like points collected during a tennis season) are used.

Each individual tennis tournament can be visualized as a tree (since there are always exactly two players competing per match). The winner of each match "rising" to the next level in a

knock-out fashion until there is only one final player left, the ultimate winner (see *figure 28*). (Each match is represented by an arrow pointing towards the winner.)

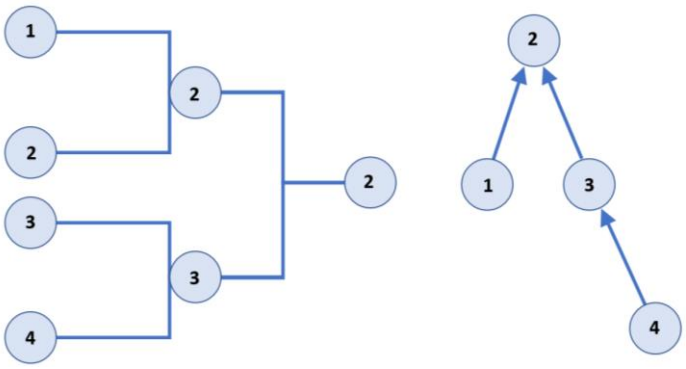


Figure 28 based on Radicchi (2011), figure 2. Tree graph to the left and directional network to the right depicting all matches and outcomes played during a single tournament with four participants.

A combined network of several tournaments has no such clear structure. Individual tournaments are linked seemingly haphazardly, creating a complex network that does not necessarily intuitively show the best player. *Figure 29* graphically represents such a setting.

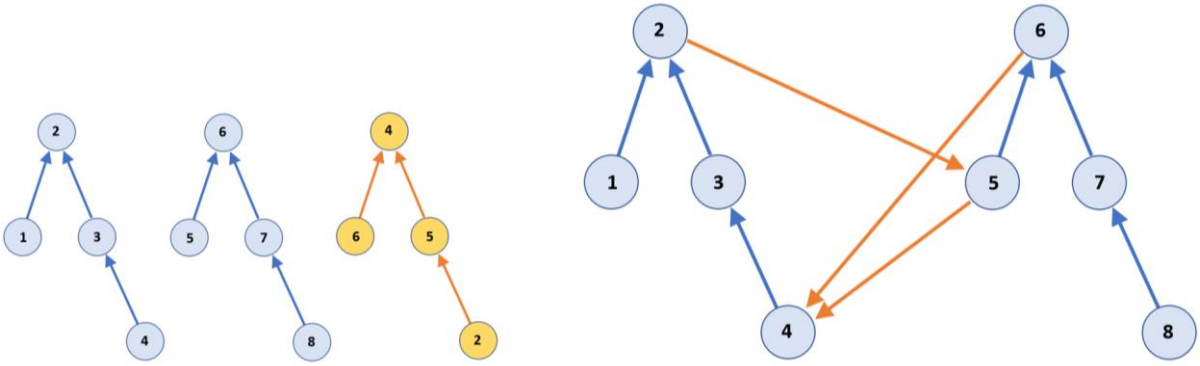


Figure 29. Eight distinct players competing in three individual tournaments (left). Combined network depicting each matches winner of the same three tournaments (right).

Considering all 133261 matches taken into account by Radicchi (2011), a weighted graph is constructed as follows: players are the nodes of the graph, arrows connect the nodes, pointing towards the winner and weighted according to the number of wins a player was able to achieve against the other.

In order to take all players into consideration at the same time, Radicchi (2011) develops the "prestige score", a measure of quality that represents the overall strength of each individual player, the highest value representing "the best player ever". This prestige has the property of flowing along the weighted arrows of the graph, representing the valuable resource that is distributed throughout the network.

In more detail, Radicchi's (2011) system is implemented as follows. We start with the following formula, taken from formula (1) of that paper:

$$P_i = (1-q)\sum_{j:\, j\neq i} P_j \frac{w_{ji}}{s_j^{out}} + \frac{q}{N} + \frac{1-q}{N}\sum_j P_j \delta(s_j^{out}). \qquad (15)$$

In this equation i and j represent tennis players and P_i represents the prestige score achieved by player i. It is calculated by solving the above system of N equations (N being the number of total players), furthermore, it is assumed that $\sum_i P_i = 1$.

We note that P_i is dependent on the score of all other players. The number of times i won against j is denoted by w_{ji}; $s_j^{out} = \sum_i w_{ji}$ denotes the "*out-strength*" of player j, representing the number of games lost by j; the term $\delta(s_j^{out})$ is used to avoid any player receiving all prestige in the case of not losing any matches (in which case he would receive prestige but it could not be redistributed to other players, effectively making such players black holes for prestige). It is 1 if $s_j^{out} = 0$ and 0 otherwise. Finally, q serves as a "*control parameter*", set at $q = 0.15$, Radicchi (2011) choosing this value "*by tradition*" in accordance with Brin & Page (1998).

Let us break down the equation into more comprehensible parts. The first summand is the most important, $\frac{w_{ji}}{s_j^{out}}$ being the proportion of all of j's losses that are due to i. If we imagine all prestige of j to be distributed among players i he lost against, then each of those players will receive a proportion of this prestige that is in accordance with the number of times he lost against i out of all his losses. It is therefore the part of the equation that factors in the "importance" of all other players. It is central to note that P_j is completely redistributed to players he lost against apart from one separate case: if $s_j^{out} = 0$, the fraction $\frac{0}{0}$ is meant to be interpreted as 0 (this detail however, is omitted by Radicchi). Such a case would occur if a given player j never loses (for example the last standing player in a single knock-out tournament). The next term, $\frac{q}{N}$ represents an equal constant distribution of prestige among all players. Finally, the last summand deals with the case of j not losing a single match. The function $\delta(s_j^{out}) = 1$ if j has 0 losses. In this case all his prestige P_j is distributed evenly among all players (including himself). Otherwise, if $\delta(s_j^{out}) = 0$, the sum is 0 and the term has no impact.

There are two distinct ways to find an (approximate) solution to equation (15). The first is to solve the system of equations (for example using Gaussian elimination), the second is to find an approximate solution using the diffusion method. In both cases, total distributed prestige is 1 (in every iteration).

Claim: Total prestige is 1 in every iteration.

Proof: To show that this requirement is fulfilled for the diffusion method, we proceed by induction. We start by distributing a total prestige of 1 to all players. Which is the case $k = 1$:

$$\sum_{i=1}^{N} P_i^{(1)} = 1. \tag{16}$$

We need to show that if we have total prestige = 1 in one step, we also have a total prestige of 1 in the following step. This process can then be repeated indefinitely. Let $P_i^{(m)}$ be the prestige of player i at the moment m (m^{th} iteration). Beginning with $\sum_{i=1}^{N} P_i^{(k-1)} = 1$, we need to show that $\sum_{i=1}^{N} P_i^{(k)} = 1$.

$$\begin{aligned}\sum_{i=1}^{N} P_i^{(k)} = \sum_{i=1}^{N}(1-q) \sum_{j:\, s_j^{out}>0}^{N} P_j^{(k-1)} \quad \frac{w_{ji}}{s_j^{out}} + \sum_{i=1}^{N} \frac{q}{N} \\ + \sum_{i=1}^{N} \frac{1-q}{N} \sum_{j:s_j^{out}=0}^{N} P_j^{(k-1)} \delta(s_j^{out})\end{aligned} \tag{17}$$

Note that in the first sum over j we excluded the indices where $s_j^{out} = 0$, so that $s_j^{out} > 0$ remains, in order to avoid division by 0. As noted above, this is assumed implicitly by Radicchi. In the rightmost sums over j we excluded the indices where $s_j^{out} > 0$, so that only $s_j^{out} = 0$ remains. This is admissible because for $s_j^{out} > 0$ it is the case that $\delta\left(s_j^{out}\right) = 0$.

We interchange the sums of the first term (the sums of the second and third term remain unchanged as they do not contain i). For $s_j^{out} = 0$ it is the case that $\delta\left(s_j^{out}\right) = 1$ and we obtain

$$\sum_{i=1}^{N} P_i^{(k)} = \sum_{j:\, s_j^{out}>0}^{N} (1-q) \sum_{i=1}^{N} P_j^{(k-1)} \quad \frac{w_{ji}}{s_j^{out}} + \sum_{i=1}^{N} \frac{q}{N} + \sum_{i=1}^{N} \frac{1-q}{N} \sum_{j:s_j^{out}=0}^{N} P_j^{(k-1)}. \tag{18}$$

We extract the constant factors of the first term (those that are not depending on i)

$$= (1-q) \sum_{j:s_j^{out}>0}^{N} P_j^{(k-1)} \frac{1}{s_j^{out}} \sum_{i=1}^{N} w_{ji} + \sum_{i=1}^{N} \frac{q}{N} + \sum_{i=1}^{N} \frac{1-q}{N} \sum_{j:s_j^{out}=0}^{N} P_j^{(k-1)}. \tag{19}$$

Because $\sum_{i=1}^{N} w_{ji} = s_j^{out}$, the first term can be further simplified to $\frac{1}{s_j^{out}} \sum_{i=1}^{N} w_{ji} = 1$. We extract further factors and receive

$$= (1-q) \sum_{j:s_j^{out}>0}^{N} P_j^{(k-1)} + \frac{q}{N} \sum_{i=1}^{N} 1 + \frac{1-q}{N} \sum_{i=1}^{N} 1 \sum_{j:s_j^{out}=0}^{N} P_j^{(k-1)}. \tag{20}$$

Because $\sum_{i=1}^{N} 1 = N$, we get

$$= (1-q) \sum_{j:s_j^{out}>0}^{N} P_j^{(k-1)} + q + (1-q) \sum_{j:s_j^{out}=0}^{N} P_j^{(k-1)}. \quad (21)$$

The conditions under the summation signs are complementary to each other and we obtain the full sum over $j = 1, \dots, N$:

$$= (1-q) \sum_{j=1}^{N} P_j^{(k-1)} + q. \quad (22)$$

As we started out under the condition that $\sum_{i=1}^{N} P_i^{(k-1)} = 1$ we have completed the induction step and have shown that total prestige of 1 is preserved through the diffusion process. For all k it is the case that

$$\sum_{j=1}^{N} P_j^{(k)} = 1. \quad (23)$$

■

In his work, Radicchi (2011) goes on to develop a separate equation for the special case of single tournaments. The graphs for such tournaments can be drawn as trees (see *figure 28*), their special feature being, that there are no cycles within the system and the number of participants shrinking by half with every level. The number of players can be denoted as $N = 2^l$, l being the total number of rounds (or matches the winner must compete in) in the tournament, because each match has two participants, one of whom drops out at every round until there is a final winner. The tree structure of such tournaments leads to the following observation: the score of any player only depends on the number of wins, in other words, two players with the same number of wins have the same score. We can show this by induction, starting at the lowest level. If we consider equation (15) and assume that a player i never wins, the first summand will be 0 (because $w_{ji} = 0$ for all j). Indeed, this player will only receive the amount given by the second and third term which distribute equal amounts of prestige to all players. In conclusion, the corresponding score will not depend on i and all players who never win, receive the same score. One level up, the first summand of equation (15) will be equal for all players losing in this round. This is due to their opponents all having had the same amount of prestige P_j as just discussed. The second and third term remain equal for all players and do not influence the ranking. The same thought-process can be repeated for all higher levels. We can further tell that additional wins lead to higher prestige scores (as long as $q \neq 1$) because the first summand of equation (15) is higher if an additional term is added, as is the case with an additional win. If the number of rounds is l, the number of ranks is $l + 1$ because there is a bottom rank for players who never win and a rank corresponding with every level of the tournament. The number of

rounds is $\log_2 N$, the number of ranks is $\log_2 N + 1$ (because half of the remaining players drop out every round and there is one additional rank for the overall winner).

Apart from arrows pointing in opposite directions, the resulting ranking is the same as would be achieved using the Points System developed by Rubinstein (1980). In both cases only the number of arrows pointing towards (resp. away from) a node is the determining factor. However, if we consider scoring, Radicchi (2011) provides a finer result.

In the case of a single (knock-out) tournament, the prestige score can be analytically calculated for any individual player, it is not necessary to solve a system of equations for all players. The solution found through Radicchi's (2011) formula (6) is exact (whereas the above universal diffusion algorithm only allows an approximation):

$$P_r = \frac{q(2-q)^r}{2^l + (2-q)^l(q-1)}. \tag{24}$$

r denotes the number of times a player wins. Due to the nature of single tournaments, as described above, the number of successful match outcomes for a player is sufficient information to rank them.

Indeed, the tree-shape also provides information about how many players will occupy each rank and how many ranks there are: because half of the players are eliminated in every round, and they all started at the same level, these eliminated players (per level) will occupy the same rank, halving the number of participants with every level. In case of a single tournament with eight participants, there will be a final ranking of one player in first place, one player in second place, two in third place and four in fourth place. There is no ranking beyond these four ranks and there is no additional network flow to influence the outcome (see *figure 30* for a possible tournament).

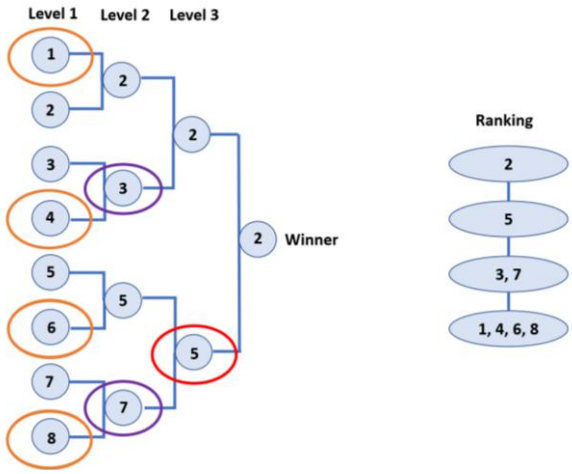


Figure 30. Single knock-out tournament with eight participants. The circled players are eliminated. All eliminated players of a common level share the same rank. There are $log_2 8 = 3$ *levels and* $log_2 8 + 1 = 4$ *ranks in this tournament.*

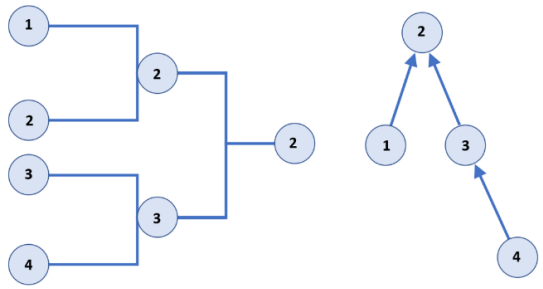
Below, we ran a simulation of Radicchi's diffusion algorithm in order to calculate an approximate solution of the system of equations for exemplary tournaments and compared the results for single tournaments with formula (24), which is formula (6) in Radicchi (2011).

5.4.1 Simulation of Radicchi's (2011) diffusion algorithm

Using Microsoft Excel Version 1807, we run a simulation of Radicchi's diffusion algorithm, building up individual components with help of the spreadsheet program. First, we simulate a single knock-out tournament with four participants. A tree-shaped graph (compare *figure 30* above) illustrates the matches and their outcomes. It is our intent to find a stable ranking. To test our result, we then calculate the exact solution, using formula (24). Next, we repeat the process with a combination of three interconnected tournaments. Again, we simulate with four players for every tournament but a total of eight different players, as one of these tournaments is interconnecting the other two. We include two cycles in this case to allow for a more complicated setup. A tree-shaped graph is no longer sufficient to illustrate the network whose participants we want to rank.

We start by building the individual components of equation (15) and implement the corresponding weighted graphs in two separate spreadsheet pages "graph representation 1" and "graph representation 2". The diffusion process is also realized in individual sheets "single tournament" and "combined tournaments".

5.4.1.1 Single tournament



		i = 1	i = 2	i = 3	i = 4
		1	2	3	4
j = 1	1	0	1	0	0
j = 2	2	0	0	0	0
j = 3	3	0	1	0	0
j = 4	4	0	0	1	0

Figure 31. Tournament between four players: tree graph, directional network, and adjacency matrix.

We use the same tournament as is depicted in *figure 29* above but transcribe the information into an adjacency-matrix, representing w_{ji} where rows are to be read as j's losses and columns represent i's wins. Row one reads as player one having no loss against himself or players three and four but losing against player 2. Column one reads as player one not achieving a single win.

We get the out-strength s_j^{out} by summing up each row separately of the adjacency-matrix to obtain the number of losses of j. A table for the division of $\frac{w_{ji}}{s_j^{out}}$ is built, to tell which proportion of j's losses is due to i. We use the if-function (Wenn-Funktion) to deal with the case of division by 0 (that is to be read as 0, see above). This function allows a differentiation between outputs. We select output 0 in case of s_j^{out}= 0 and $\frac{w_{ji}}{s_j^{out}}$ otherwise. The function $\delta(s_j^{out})$ is implemented by again using the if-function to provide the output 1 if s_j^{out}= 0 and 0 otherwise. For visual documentation of these processes, please refer to Appendix 8.3.

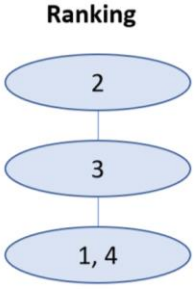
With these building-blocks, we can assemble equation (15) using the sumproduct-function (Summenprodukt-Funktion). This function allows the scalar multiplication of two vectors. We start with a uniform distribution of prestige (although any start vector with $\sum = 1$ could be used). We choose 0.25 as our start vector and allow for variable q but select $q = 0.15$ as default value. (The implementation is realized through the formula =(1-$D$3)*SUMMENPRODUKT(C$6:C$9;'graph representation 1'!$Q$8:$Q$11)+$D$3/4 +((1-$D$3)/4)*SUMMENPRODUKT(C$6:C$9;'graph representation 1'!$O$8:$O$11). The cells represent the following: $D$3 = q, C$6:C$9 represents the start vector 0.25 for all P_i, $Q$8:$Q$11 represents the first column of the $\frac{w_{ji}}{s_j^{out}}$-table, $O$8:$O$11 represents the function $\delta(s_j^{out})$.) We find that the diffusion process quickly leads to stable values. By using the Microsoft Excel feature of "shifting" the formula (thereby copying cells held by the $-symbol but equally shifting other cells that are linked), we run 30 iterations of the above and arrive at a stable result (for eight positions after the decimal point) for all P_i after 22 iterations. The value used for P_j in any given iteration is the output of the previous run. It is noteworthy that the two participants losing in the first round, P_1 and P_4, indeed have identical results in all iterations.

Diffusion process

Control parameter q: 0,15

	Start vector	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6
P1	0,25	0,090625	0,14707031	0,13987354	0,13426605	0,13914074	0,13704597
P2	0,25	0,515625	0,48175781	0,45536963	0,47830936	0,46845165	0,47059245
P3	0,25	0,303125	0,22410156	0,2648833	0,25315855	0,25326688	0,2553156
P4	0,25	0,090625	0,14707031	0,13987354	0,13426605	0,13914074	0,13704597
Prestige sum	1	1	1	1	1	1	1

Iteration 22	Iteration 23	Iteration 24	Iteration 25	Iteration 26	Iteration 27	Iteration 28	Iteration 29	Iteration 30
0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043
0,47060846	0,47060846	0,47060846	0,47060846	0,47060846	0,47060846	0,47060846	0,47060846	0,47060846
0,25438295	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295
0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043
1	1	1	1	1	1	1	1	1



We find the ranking of player 2 in first place, player 3 in second place and players 1 and 4 sharing third position.

The calculated ranking corresponds with the intuitive ranking apparent from the graph. To test our result, we compare our findings to equation (24) for this simulation.
For the exact solution, we need the number of wins r, the number of rounds $l = 2$ and q (selected at $q = 0.15$ following Radicchi's (2011) example). We create a table for r and calculate the results for P_i. (We use the following formula: =(D$3*(2-D$3)^C16)/(2^2+(2-D$3)^2*(D$3-1)). The cells represent the following: D$3 = q, C16 = r.) Comparing the exact solution with the result of the final iteration of our simulation, we find them to be identical.

Exact solution (Radicchi (2011), equation (6)) **Diffusion process**

	r (number of wins)		**Iteration 30**
P1	0	0,1375043	0,1375043
P2	2	0,47060846	0,47060846
P3	1	0,25438295	0,25438295
P4	0	0,1375043	0,1375043
	Prestige sum	1	

The induced ranking is independent of the value of q for $0 \leq q < 1$. This is due to the fact that equation (24) is increasing in the variable r. For $q = 1$ prestige would be equally divided between all players, which is apparent from both formulas (15) and (24). If $q = 1$ formula (15) would only consist of the second term $\frac{1}{N}$ and formula (24) read $\frac{1}{2^l} = \frac{1}{N}$. We created a graph using different values for q $(0, 0.15, 0.3, 0.5, 0.7, 0.85, 1)$, showing that the lines converge but do not cross apart from merging at $q = 1$. P_1 and P_4 are identical because each of them has 0 wins and they therefor achieve the same score.

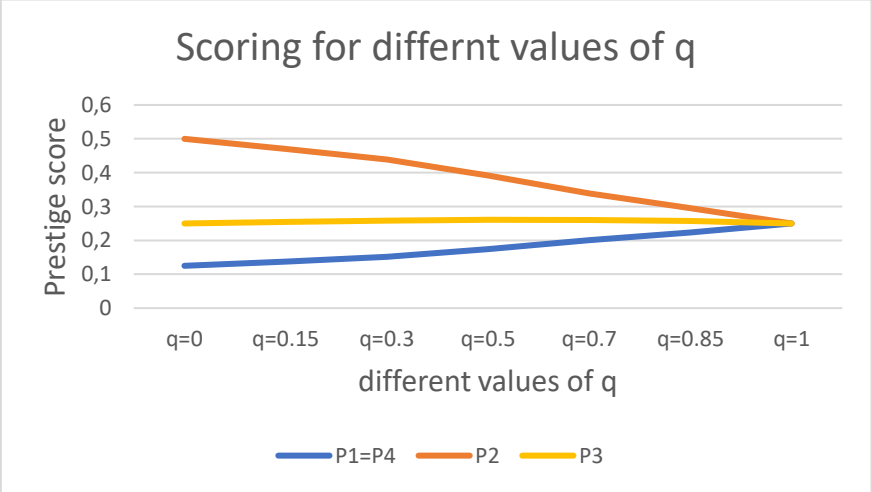
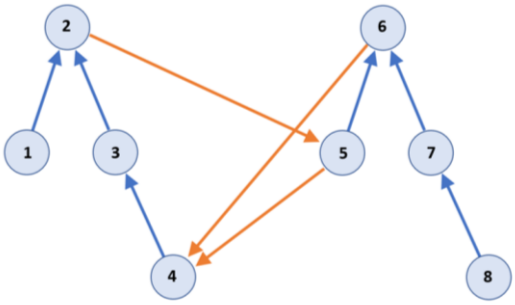


Figure 32. Scoring for different values of q. P1 and P4 are identical. All lines converge towards the equal score of 0.25 at q=1.

5.4.1.2 Combined tournaments



		i = 1	i = 2	i = 3	i = 4	i = 5	i = 6	i = 7	i = 8
		1	2	3	4	5	6	7	8
j = 1	1	0	1	0	0	0	0	0	0
j = 2	2	0	0	0	0	1	0	0	0
j = 3	3	0	1	0	0	0	0	0	0
j = 4	4	0	0	1	0	0	0	0	0
j = 5	5	0	0	0	1	0	1	0	0
j = 6	6	0	0	0	1	0	0	0	0
j = 7	7	0	0	0	0	0	1	0	0
j = 8	8	0	0	0	0	0	0	1	0

Figure 33. Combined tournament between eight players in total: directional network, and adjacency matrix

Again, we use a tournament depicted above (*figure 33*) and transcribe the information into an adjacency-matrix representing w_{ji} where rows are to be read as j's losses and columns represent i's wins. In this case, the tournament between players 2, 4, 5 and 6 (orange) connects the two single tournaments between players 1, 2, 3, 4 and 5, 6, 7, 8 respectively (both blue). A cycle between players $2 \rightarrow 5 \rightarrow 4 \rightarrow 3 \rightarrow 2$ adds interest to the network structure as this means that the best player of one single tournament (player 2) loses to one of the worst player of another tournament (player 5). A second cycle can be found (in shape of a figure eight loop) between players $2 \rightarrow 5 \rightarrow 6 \rightarrow 4 \rightarrow 3 \rightarrow 2$.

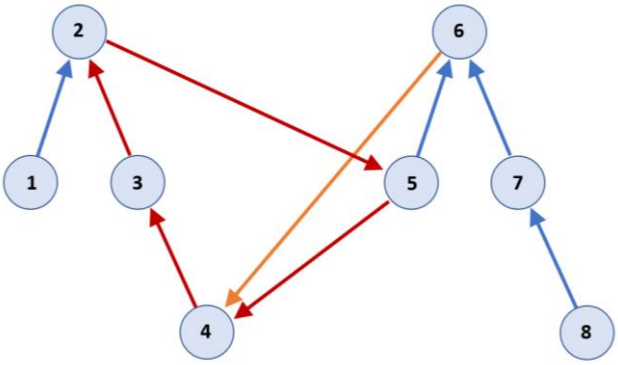


Figure 34. Cycles in combined tournament from figure 33: 2-5-4-3-2 and 2-5-6-4-3-2.

The idea of finding a reasonable ranking that is dependent on the opponents' qualities now gains interest. We can see that players 1 and 8 never win, players 3, 5 and 7 all achieve one win, players 2, 4 and 6 win twice. Player 5 on the other hand loses twice, once to player 4 and once to player 6. We employ the same methods used above for single tournaments to get the out-strength s_j^{out}, a table for $\frac{w_{ji}}{s_j^{out}}$ and the function $\delta(s_j^{out})$. The fact that player 5 loses twice leads

to $s_5^{out} = 2$ as well as $\frac{w_{5,4}}{s_5^{out}} = \frac{w_{5,6}}{s_5^{out}} = 0.5$. We can see that all players lose at least once, leading to all $\delta(s_j^{out})$ being 0 and the third term of equation (15) dropping out.

For a visual documentation of the simulation process, please refer to Appendix 8.3.

Nevertheless, we assemble equation (15) in the same way we did for the single tournament and run the diffusion algorithm. We use an even distribution of prestige at the outset and select 0.125 as a start vector, $q = 0.15$ remains unchanged. We find that the diffusion process does not converge as quickly towards stable values in our simulation as it did in the single tournament experiment. Indeed, we run 100 iterations and it is only after 94 iterations that we arrive at constant values for eight places after the decimal point. We arrive at the following ranking:

Diffusion process

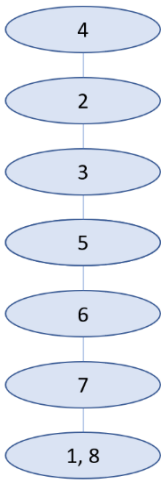
Control parameter q

0,15

	Start vector	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6
P1	0,125	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875
P2	0,125	0,23125	0,1409375	0,17932031	0,2119457	0,23967729	0,1925336
P3	0,125	0,125	0,17015625	0,20853906	0,24116445	0,18570129	0,17391537
P4	0,125	0,178125	0,22328125	0,26166406	0,19641328	0,18254749	0,20611933
P5	0,125	0,125	0,2153125	0,13854688	0,17117227	0,19890385	0,22247569
P6	0,125	0,178125	0,178125	0,13974219	0,1071168	0,12098259	0,13276851
P7	0,125	0,125	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875
P8	0,125	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875
Prestige sum	1	1	1	1	1	1	1

Iteration 93	Iteration 94	Iteration 95	Iteration 96	Iteration 97	Iteration 98	Iteration 99	Iteration 100
0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875
0,20196502	0,20196502	0,20196502	0,20196502	0,20196502	0,20196502	0,20196502	0,20196502
0,19679708	0,19679708	0,19679708	0,19679708	0,19679708	0,19679708	0,19679708	0,19679708
0,20946715	0,20946715	0,20946715	0,20946715	0,20946715	0,20946715	0,20946715	0,20946715
0,19042026	0,19042026	0,19042026	0,19042026	0,19042026	0,19042026	0,19042026	0,19042026
0,12916299	0,12916299	0,12916299	0,12916299	0,12916299	0,12916299	0,12916299	0,12916299
0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875
0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875
1	1	1	1	1	1	1	1

Ranking



Of course, equation (24) is not applicable in this more general case and cannot help us discern an exact solution. Other methods, like the Gaussian elimination could be utilized to find an exact solution but such methods can get ungainly with increasing network size.

We also ran the simulation for several values of the control parameter q and noticed that the induced ranking is not independent of this parameter. We again created a graph using the different values for q $(0, 0.15, 0.3, 0.5, 0.7, 0.85, 1)$ we employed and arrived at the interesting observation that the lines showing different scoring for different values of q intersect: player 6

"overtakes" players 3 and 5 around $q = 0.5$, player 2 passes player 4 between $q = 0.5$ and $q = 0.7$, thereby changing the overall ranking. It remains unchanged that all lines converge to the uniform distribution, in this case 0.125, for $q = 1$.

We compared our results to a slightly modified Points System (reversing arrows to count in-going edges instead of the opposite) and found that it yields the following ranking that does not take into account the opponents' strength:

This is similar to the ranking the diffusion method yields for higher values of q (after the final intersection of lines in our graph), where players 2, 4 and 6 occupy the first three positions, players 3, 5 and 7 the following three positions and players 1 and 8 sharing the bottommost rank.

6. Conclusion

In this work, we have investigated four distinct measures of centrality: degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. While some parts of this work are technical, this did not prevent us from delving into the subject matter. The mathematics involved in the studied literature may rise beyond a basic level at times but the increase in understanding achieved is rewarding and egalizes the effort spent.

Social network analysis is an important tool in modern economics. To establish a ranking of actors, measures of centrality can be valuable instruments. Used correctly, they can help to identify potentially influential individuals, who have the power to reach large network areas. High closeness centrality values could indicate such a situation. Whenever it is an objective to achieve extensive broadcasting, be it of information, ideas or technology, this measure is useful. Imagine a new vaccine that can prevent the common cold on the market. Valuable vendors of this medicine would be people with far and wide reach into the network of doctors, who would in turn administer the drug. High values of betweenness centrality indicate individuals, who have the power to control the pace with which goods flow between nodes. Economic value is achieved by actors, who can make use of such a brokerage position. These may be individuals who demand compensation for their work of simple passing on information that would otherwise have to take a longer route. A journalist may be tempted to use such individuals in order to be the first to report breaking news. Also, advantages might be gained from knowing actors, who in turn are in contact with other powerful players. Consider a CEO, who shares friendships with powerful business leaders. Eigenvector centrality could help to identify this person as a target to be befriended.

Centrality measures can be applied to a host of situations, whenever basic network structures arise. Equally, rankings can be derived from the obtained measures, but trust in these rankings must be preceded by the understanding of the context. It is remarkably simple to use a wrong measure and not realize it, devoiding any work involved. It seems prudent to use caution both when applying a measure of centrality and when accepting results based on such measures. In any case, interpretation must correlate with the respective measure.

If a ranking is for instance based on the points system, it should not be used to measure a nodes' brokering power as this depends on further reach and access within the network. Closeness centrality and betweenness centrality would better fit these requirements. On the other hand, to measure and rank the popularity of an actor, degree centrality seems best suited. Betweenness centrality may convey the wrong picture if a node with just two acquaintances bridges otherwise

separate clusters of a network. All flow between clusters would travel through this node, resulting in a large betweenness centrality.

Powerful influencers in a network can be identified by ranking actors through either closeness or betweenness centralities and eigenvector centrality is the method of choice if one node's importance depends on the importance of its neighbours, as in the evaluation of a chess players' strength.

We have established that the appropriate measure of centrality used in the correct way can yield significant assessments of network positions that can be translated into powerful rankings. Further research into obtaining these positions would be of great interest.

7. References

AmericanMathematicalSociety. (2018a). *MathSciNet - Mathematical Reviews*. Abgerufen am 10. Sep 2018 von https://mathscinet.ams.org/mathscinet/

AmericanMathematicalSociety. (2018b). *MathSciNet - Mathematical Reviews*. Abgerufen am 10. Sep 2018 von Citation Database Help Topics - Journal information returned from the Citation Database: https://mathscinet.ams.org/mathscinet/help/citation_database_help_full.html

Bavelas, A. (1950). Communication patterns in task-oriented groups}. *The Journal of the Acoustical Society of America*, *22*(6), S. 725-730.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral science, 10*(2), S. 161-163.

Bernard, H. R., Johnsen, E., Killworth, P., & Robinson, S. (1989). Estimating the Size of an Average Personal Network and of an Event Subpopulation. In M. (. Kochen, *The Small World* (S. 159-75). Norwood, NJ: Ablex Publishing Co.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology, 2*(1), S. 113-120.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology, 92*(5), S. 1170-1182.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks, 27*(1), S. 55-71.

Bourdieu, P. (2011). The forms of capital. *Cultural theory: An anthology, 1*, S. 81-93.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems, 30*(1-7), S. 107-117.

Burt, R. S. (1982). Toward a structural theory of action: network models of social Structure, Perception, and Action.

Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition.* Harvard University Press.

Burt, R. S. (2000). The Network Structure of Social Capital. *Research in Organizational Behavior*(22), S. 345-423.

Butts, C. T. (2008). Social network analysis with sna. *Journal of statistical software, 24*(6), S. 1-51.

Chartrand, G. (1984). *Introductory Graph Theory.* Dover Publications Inc.

Chung, K., Hossain, L., & Davis, J. (2005). *Exploring sociocentric and egocentric approaches for social network analysis.* Proceedings of the 2nd international conference on knowledge management in Asia Pacific.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*, S. 95-120.

Connelly, B., Tihanyi, L., Crook, T., & Gangloff, K. (2014). Tournament theory: Thirty years of contests and competitions. *Journal of Management, 40*(1), S. 16-47.

Czepiel, J. A. (1974). Word of mouth processes in the diffusion of a major technological innovation. *Journal of Marketing Research, 11*, S. 172-180.

Davis, J. A. (1970). Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *American Sociological Review*, S. 843-851.

Duch, J., Waitzman, J., & Amaral, L. (2010). Quantifying the performance of individual players in a team activity. *PloS one, 5*(6), S. e10937.

Eppstein, D., & Wang, J. (2004). Fast Approximation of centrality. *Journal of Gaph Algorithms and Applications, 8*(1), S. 39-45.

Feld, S., & Carter, W. (1998). Foci of Activities As Changing Contexts for Friendship. In R. Adams, & A. Graham (Hrsg.), *Placing friendship in context* (S. 136-152). Cambridge, UK: Cambridge University Press.

Franks, H., Griffith, H., & Anand, S. (2014). Learning agent influence in MAS with complex social networks. *Autonomous Agents and Multi-Agent Systems, 28*(5), S. 836-866.

Fredman, M., & Tarjan, R. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. *Journal for the Association of Computing Machinery, 34*, S. 596-615.
Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, S. 35-41.
Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), S. 215-239.
Frobenius, G. (1912). Über Matrizen aus nicht negativen Elementen. *Sitzung der physikalisch-mathematischen Classe*, S. 456-477.
Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology, 78*(6), S. 1360-1380.
Granovetter, M. (1982). The strength of weak ties: A network theory revisited. In P. Marsden, & N. Linn (Hrsg.), *Social Structure and Network Analysis* (S. 105-130). Beverly Hills, Ca: Sage.
Hanneman, R., & Riddle, M. (2005). *Introduction to social network methods.* Riverside, CA: University of California, Riverside. Abgerufen am 14. Jul 2018 von http://faculty.ucr.edu/~hanneman/
Holland, P., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative group studies, 2*(2), S. 107-124.
Johnsen, E., Killworth, P., & Robinson, S. (1989). Estimating the size of an average personal network and of an event subpopulation. In *The small world* (S. 159-175). Norwood, NJ: Ablex Publishing Co.
Johnson, D. (1977). Efficient algorithms for shortest paths in sparce networks. *Journal of the Assiciation for Computing Machinery, 24*, S. 1-13.
Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings.* OUP USA.
Kono, C., Palmer, D., Friedland, R., & Zafonte, M. (1998). Lost in space: The geography of corporate interlocking directorates. *American journal of sociology, 103*(4), S. 863-911.
Landau, H. G. (1953). On dominance relations and the structure of animal societies: III The condition for a score structure. *Bulletin of Mathematical Biology, 15*(2), S. 143-148.
Lazarsfeld, P., & Merton, R. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society, 18*(1), S. 18-66.
Lazear, E., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of political Economy, 89*(5), S. 841-864.
Leader, I. (2001). Friends and strangers. *Millenium Mathematics Project, University of Cambridge*.
Levin, D., & Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management science, 50*(11), S. 1477-1490.
McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *MC medical research methodology, 7*(1), S. 30.
Mitchell, J. (Hrsg.). (1969). *Social networks in urban situations: analyses of personal relationships in Central African towns.* Manchester: University Press.
Müller, M. (Hrsg.). (2017). *Überraschende Mathematische Kurzgeschichten: Ausgewählte Artikel des jungen Ablegers der Zeitschrift "Die Wurzel".* Springer-Verlag.
Nieminen, J. (1974). On the centrality in a graph. *Scandinavian journal of psychology, 15*(1), S. 332-336.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford digital library technologies project*.
Radicchi, F. (2011). Who is the best player ever? A complex network analysis of the history of professional tennis. *PloS one, 6*(2), S. e17249.
Ramsey, F. P. (1930). On a problem of formal logic. *Proceedings of the London Mathematical Society, 30*, S. 264-286.
Rogers, D. L. (1974). Sociometric analysis of interorganizational relations: application of theory and measurement. *Rural Sociology, 39*(4), S. 487.

Rubinstein, A. (1980). Ranking the participants in a tournament. *SIAM Journal of Applied Mathematics, 38*(1), S. 108-111.

Ryder, T., McDonald, D., Blake, J., Parker, P., & Loiselle, B. (2008). Social networks in the lek-mating wire-tailed manakin (Pipra filicauda). *Proceedings of the Royal Society of London B: Biological Sciences, 275*(1641), S. 1367-1374.

Scott, J. (1997). *Social Network Analysis: A Handbook.* Sage Publications.

Simmel, G. (1950). *The Sociology of Georg Simmel.* (K. H. Wolff, Übers.) Glencoe, Illinois : The Free Press. Abgerufen am Sep 2018 von https://archive.org/stream/sociologyofgeorg030082mbp/sociologyofgeorg030082mbp_djvu.txt

Sloane, N. J. (2018). *The On-Line Encyclopedia of Integer Sequences*. Abgerufen am 15. Sep 2018 von https://oeis.org/

Travers, J., & Milgram, S. (1967). The small world problem. *Phychology Today, 1*(1), S. 61-67.

Vernon, R., & Hoover, E. (1962). Anatomy of a Metropolis. In C. Kadushin, *Understanding social networks: Theories, concepts, and findings.* OUP USA.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Bd. 8). Cambridge: University Press.

8. Appendix

8.1 Abstract

A social network is a network of individuals (actors) who are connected through a common relationship. To evaluate actors in social networks, we consider such networks from a graph-theoretical perspective. There are two possibilities to visualize such a network. One is the sociogram, where actors are depicted as nodes and relationships as ties between them. The other is by means of an adjacency matrix, where both rows and columns indicate actors and entries represent the corresponding relationship. If we consider a social network, the entry „1“ may indicate friendship and „0“ the absence thereof. Also, weighted connections are possible.

In this text we will describe in detail four mathematical methods that allow the formal implementation of the intuitive notion of „centrality“. The considered measures of centrality are degree-, betweenness-, closeness-, and eigenvector centrality. There are explicit formulas to calculate the valuation of individual actors for the first three of these measures. Degree centrality counts the number of incoming edges, betweenness centrality is concerned with the number of shortest paths an actor lies on, and the related notion of closeness centrality measures an actor’s mean distance to all other actors. Finally, eigenvector centrality does not allow the calculation of individual values alone but needs to be calculated (possibly approximately) by means of a system of equations. This can be done through matrix calculus.

Because of their versatility, the introduced measures of centrality are of great economic interest. Famously, Google’s PageRank-algorithm is a variant of eigenvector centrality. Here, an item’s ranking determines its further exposure. This visibility is of economic relevance for businesses, since users usually will not look past the first few entries. Other potential applications are the estimation of a Twitter users’ popularity by counting his followers (degree centrality), which can be of interest when placing advertisements. The detection of influential actors through either betweenness- or closeness centrality, which can be used to identify brokers, is another field of application.

Four scientific articles are further analyzed in depth, providing links from each considered measure of centrality to contemporary topics. This work is rounded off with two mathematical proofs and a self-devised simulation of an algorithm of the PageRank type.

8.2 Zusammenfassung

Ein soziales Netzwerk besteht aus Akteuren, die durch eine gemeinsame Beziehung verbunden sind. Um Individuen in einem sozialen Netzwerk bewerten zu können, betrachten wir derartige Netzwerke aus einem graphentheoretischen Blickpunkt. Mögliche Darstellungsformen sind einerseits das Soziogramm, bei dem Individuen als Knoten und Beziehungen als Verbindungen dargestellt werden, sowie andererseits die Darstellung als Nachbarschaftsmatrix, bei der Zeilen und Spalten Individuen bezeichnen und Einträge in der Matrix die Qualität der Verbindung beschreiben. Wird als Beziehung etwa das Bestehen einer Freundschaft gewählt, so kann dies beispielsweise durch die Einträge „1“ oder „0“ in der Matrix realisiert werden. Auch gewichtete Verbindungen sind möglich.

Vier mathematische Methoden, die den intuitiven Begriff der Zentralität in einem Netzwerk formal beschreiben, werden genau erläutert. Die betrachteten Zentralitätsbegriffe lauten Degree-, Betweenness-, Closeness-, und Eigenvector Centrality. Die ersten drei dieser Maße können durch explizite Formeln berechnet werden, wodurch die individuelle Bewertung eines einzelnen Knotens leicht möglich ist. Degree Centrality beschäftigt sich mit der Anzahl eingehender Verbindungen, Betweenness Centrality bestimmt die Anzahl an kürzesten Pfaden, auf denen ein Knoten liegt, der verwandt Begriff der Closeness Centrality bestimmt den mittleren Abstand eines Knotens zu allen anderen. Eigenvektorzentralität kann nicht für individuelle Knoten berechnet werden, hier ist die (mitunter näherungsweise) Lösung eines Gleichungssystems nötig für das sich Matrizenrechnung anbietet.

Die vorgestellten Maße haben große ökonomische Bedeutung, da sie auf verschiedenste Netzwerksituationen angewandt werden können. Große Bekanntheit hat der Google PageRank Algorithmus, nach dem die Ergebnisse der Suchmaschine Google geordnet wurden. Die erhaltene Reihung bestimmt die für Unternehmen wirtschaftlich bedeutsame Sichtbarkeit und basiert auf einer Variante der Eigenvector Centrality. Andere Anwendungsmöglichkeiten sind beispielsweise das für Werbezwecke bedeutsame Bewerten der Beliebtheit eines Twitter-Users durch Zählen seiner Follower (Degree Centrality). Die Identifizierung einflussreicher Individuen, die etwa Vermittlerrollen übernehmen können, kann durch Betweenness- oder Closeness Centrality erfolgen.

Um einen Bezug zwischen Zentralitätsmaßen und aktueller Forschung herzustellen, werden vier wissenschaftliche Artikel im Detail analysiert. Zwei mathematische Beweise und eine selbsterstellte Simulation eines Algorithmus der PageRank-Art, runden die Arbeit ab.

Graph representation for a single tournament with 4 participants

		i = 1	i = 2	i = 3	i = 4
		1	2	3	4
j = 1	1	0	1	0	0
j = 2	2	0	0	0	0
j = 3	3	0	1	0	0
j = 4	4	0	0	1	0

sj out (total number of losses of j)

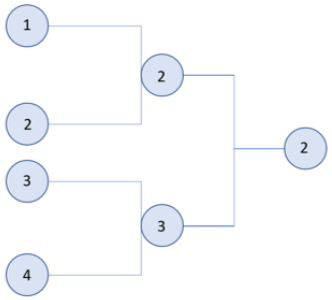
1
0
1
1

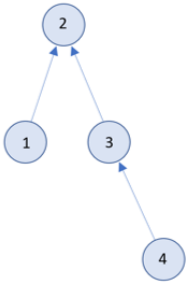
δ(sj out) equals 1, if sj out 0, and 0 otherwise.

0
1
0
0

Table: wji/sj out
Which proportion of all losses of j is due to i

0	1	0	0
0	0	0	0
0	1	0	0
0	0	1	0





w_{ji} - number of i's (column) wins against j (row).
Column i contains the numbers of i's wins;
Row j contains the number of j's losses.

In equation (1), Radicchi divides by 0, but the expression $0/0$ is to be interpreted as 0.

Control parameter q

0,15

	Start vector	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 21	Iteration 22	Iteration 23	Iteration 24	Iteration 25	Iteration 26	Iteration 27	Iteration 28	Iteration 29	Iteration 30
P1	0,25	0,090625	0,14707031	0,13987354	0,13426605	0,13914074	0,137504298	0,137504296	0,137504297	0,137504297	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043
P2	0,25	0,515625	0,48175781	0,45536963	0,47830936	0,46845165	0,470608453	0,470608458	0,470608456	0,470608456	0,47060846	0,47060846	0,47060846	0,47060846	0,47060846	0,47060846
P3	0,25	0,303125	0,22410156	0,2648833	0,25315855	0,25326688	0,254382951	0,25438295	0,254382949	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295	0,25438295
P4	0,25	0,090625	0,14707031	0,13987354	0,13426605	0,13914074	0,137504298	0,137504296	0,137504297	0,137504297	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043	0,1375043

Prestige sum	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Exact solution (Radicchi (2011), equation (6)) Diffusion process

	r (number of wins)	Iteration 30
P1	0	0,1375043
P2	2	0,47060846
P3	1	0,25438295
P4	0	0,1375043

Prestige sum	1

l (number of rounds in the tournament) = 2

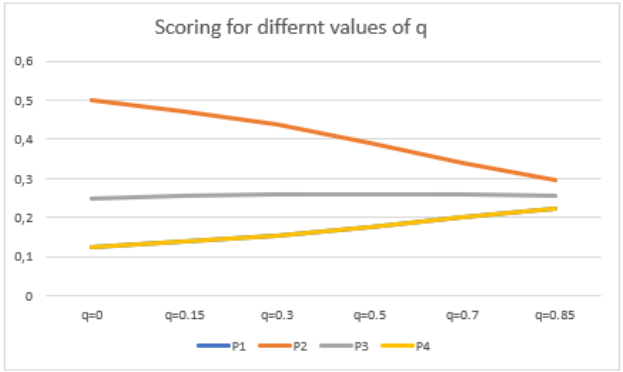
The implemented diffusion process quickly leads to stable values. The calculated ranking corresponds with the intuitive ranking apparent from the graph and also with the exact solution found through Radicchi's single tournament equation.

Prestige sum is 1 for all iterations.
P1 and P4 need to be identical in all iterations (if start values are the same, otherwise after the first iteration) due to the fact that both players only receive the constant prestige value all players receive and cannot collect any prestige through victory.

The control parameter q was chosen 0,15 in correspondence with Radicchi (2011)

Iteration 30

	q=0	q=0.15	q=0.3	q=0.5	q=0.7	q=0.85
P1	0,125	0,1375043	0,15174507	0,17391304	0,2004008	0,2235886
P2	0,5	0,47060846	0,43854325	0,39130435	0,33867735	0,29569592
P3	0,25	0,25438295	0,25796662	0,26086957	0,26052104	0,25712689
P4	0,125	0,1375043	0,15174507	0,17391304	0,2004008	0,2235886



Graph representation for three combined tournaments with 8 participants

		i = 1	i = 2	i = 3	i = 4	i = 5	i = 6	i = 7	i = 8
		1	2	3	4	5	6	7	8
j = 1	1	0	1	0	0	0	0	0	0
j = 2	2	0	0	0	0	1	0	0	0
j = 3	3	0	1	0	0	0	0	0	0
j = 4	4	0	0	1	0	0	0	0	0
j = 5	5	0	0	0	1	0	1	0	0
j = 6	6	0	0	0	1	0	0	0	0
j = 7	7	0	0	0	0	0	1	0	0
j = 8	8	0	0	0	0	0	0	1	0

sj out (total number of losses of j)

1
1
1
1
2
1
1
1

δ(sj out) equals 1, if sj out 0, and 0 otherwise.

0
0
0
0
0
0
0
0

Table: wji/sj out
Which proportion of all losses of j is due to i

0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0,5	0	0,5	0	0
0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0

w_{ji} - number of i's (column) wins against j (row).
Column i contains the numbers of i's wins;
Row j contains the number of j's losses.

In equation (1), Radicchi divides by 0, but the expression $0/0$ is to be interpreted as 0.

Diffusion process

Control parameter q

0,15

	Start vector	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6	Iteration 7	Iteration 8	Iteration 94	Iteration 95	Iteration 96	Iteration 97	Iteration 98	Iteration 99	Iteration 100
P1	0,125	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875
P2	0,125	0,23125	0,1409375	0,1793203	0,2119457	0,2396773	0,1925336	0,1825156	0,1995462	0,201965	0,201965	0,201965	0,201965	0,201965	0,201965	0,201965
P3	0,125	0,125	0,1701563	0,2085391	0,2411645	0,1857013	0,1739154	0,1939514	0,2109821	0,1967971	0,1967971	0,1967971	0,1967971	0,1967971	0,1967971	0,1967971
P4	0,125	0,178125	0,2232813	0,2616641	0,1964133	0,1825475	0,2061193	0,2261554	0,2176401	0,2094672	0,2094672	0,2094672	0,2094672	0,2094672	0,2094672	0,2094672
P5	0,125	0,125	0,2153125	0,1385469	0,1711723	0,1989038	0,2224757	0,1824036	0,1738882	0,1904203	0,1904203	0,1904203	0,1904203	0,1904203	0,1904203	0,1904203
P6	0,125	0,178125	0,178125	0,1397422	0,1071168	0,1209826	0,1327685	0,1427865	0,1257559	0,129163	0,129163	0,129163	0,129163	0,129163	0,129163	0,129163
P7	0,125	0,125	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875	0,0346875
P8	0,125	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875	0,01875

Prestige su	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Iteration 100

	q=0	q=0.15	q=0.3	q=0.5	q=0.7	q=0.85	q=1
P1	0	0,01875	0,0375	0,0625	0,0875	0,10625	0,125
P2	0,222463816	0,201965	0,1852924	0,1680328	0,1535027	0,1411834	0,125
P3	0,22208536	0,1967971	0,173632	0,1485656	0,132509	0,1266391	0,125
P4	0,221961122	0,2094672	0,1944743	0,1721311	0,1500299	0,1359273	0,125
P5	0,222467791	0,1904203	0,1672047	0,1465164	0,1335508	0,1274275	0,125
P6	0,11102191	0,129163	0,1406466	0,1460041	0,1416576	0,1341352	0,125
P7	0	0,0346875	0,06375	0,09375	0,11375	0,1221875	0,125
P8	0	0,01875	0,0375	0,0625	0,0875	0,10625	0,125

The implemented diffusion process does not quickly lead to stable values in our experiment.

Prestige sum is 1 for all iterations.

The control parameter q was chosen 0.15 in correspondence with Radicchi (2011), who notes that the original PageRank algorithm also employed this value.

Additional diffusion processes were calculated for $q = 0; q = 0.3; q = 0.5; q = 0.7; q = 0.85$ and $q = 1$. The graph visualizes the fact that different values of q can influence the ranking derived from the prestige values, if we consider combined tournaments. A careful selection of q is therefore necessary.