Graph Theory

Dr. Jens Dörpinghaus

doerpinghaus@uni-koblenz.de

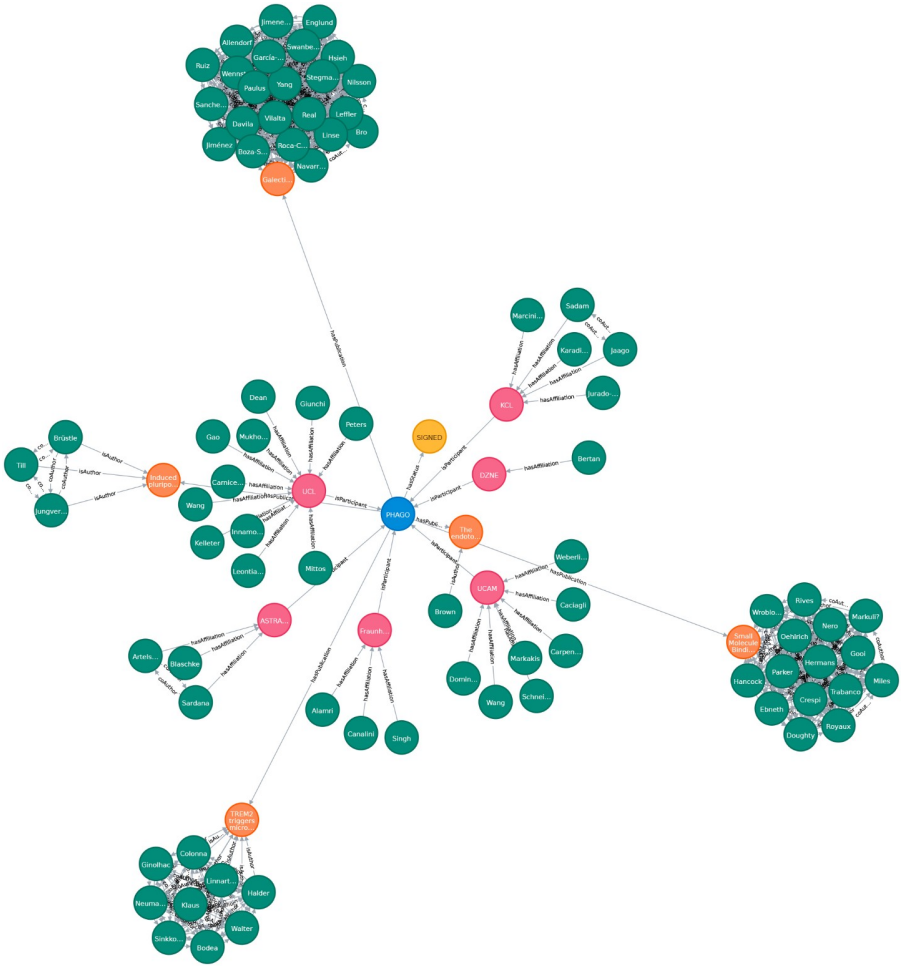
Mathematisches Institut
Universität Koblenz-Landau
Campus Koblenz

Federal Institute for Vocational Education and Training (BIBB)
Robert-Schuman-Platz 3
53175 Bonn

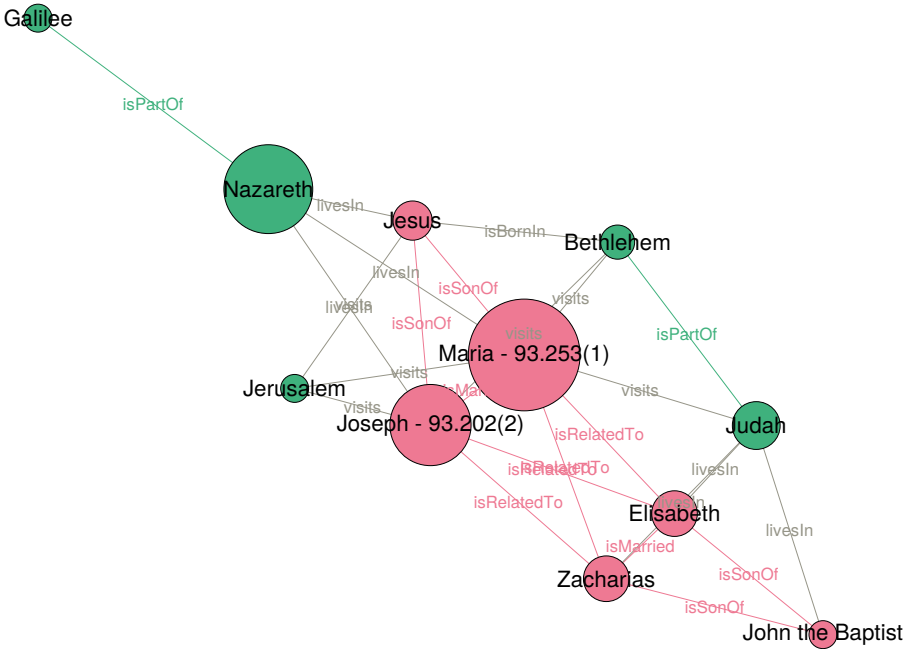
Summer 2022

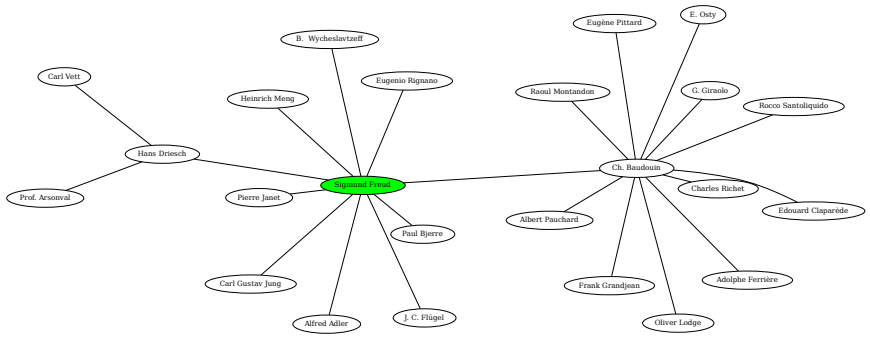
6 Random Graphs and Social Networks

- Introduction
- Social Networks
- Network Measures
- Strong and weak ties
- Homophily
- Community Detection
- Network Dynamics
- Random graphs



- Social network analysis (SNA) is an emerging topic and is increasingly becoming a vital factor in other scientific domains.
- They play an important role in the social sciences and have been widely used for several decades, both in theory and in application.
- It is an important issue to understand social interactions and networks and how they influence society.
- In the last few years there has been a growing interest in using social networks in historical sciences.
- Quite recently, considerable attention has been paid to social networks in religious studies.


Galilee
isPartOf
Nazareth
livesIn
Jesus
isBornIn
Bethlehem
livesIn
isSonOf
visits
livesIn
isSonOf
visits
isPartOf
Maria - 93.253(1)
Jerusalem
visits
visits
Joseph - 93.202(2)
isRelatedTo
Judah
isRelatedTo
livesIn
isRelatedTo
Elisabeth
livesIn
isMarried
Zacharias
isSonOf
isSonOf
John the Baptist


Carl Vett
B. Wycheslavtzeff
Eugène Pittard
E. Osty
Heinrich Meng
Eugenio Rignano
Raoul Montandon
G. Giraolo
Rocco Santoliquido
Hans Driesch
Ch. Baudouin
Prof. Arsonval
Sigmund Freud
Pierre Janet
Charles Richet
Edouard Claparède
Albert Pauchard
Paul Bjerre
Carl Gustav Jung
Frank Grandjean
Adolphe Ferrière
Alfred Adler
J. C. Flügel
Oliver Lodge

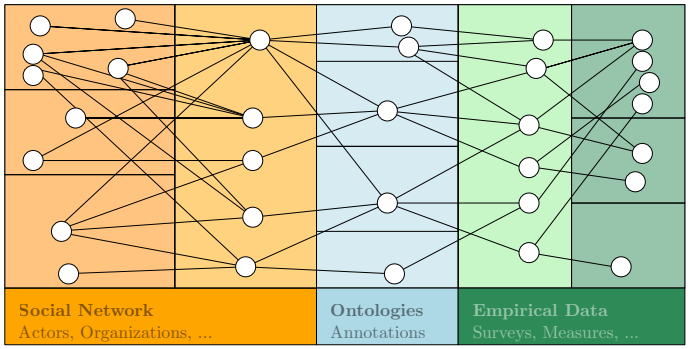
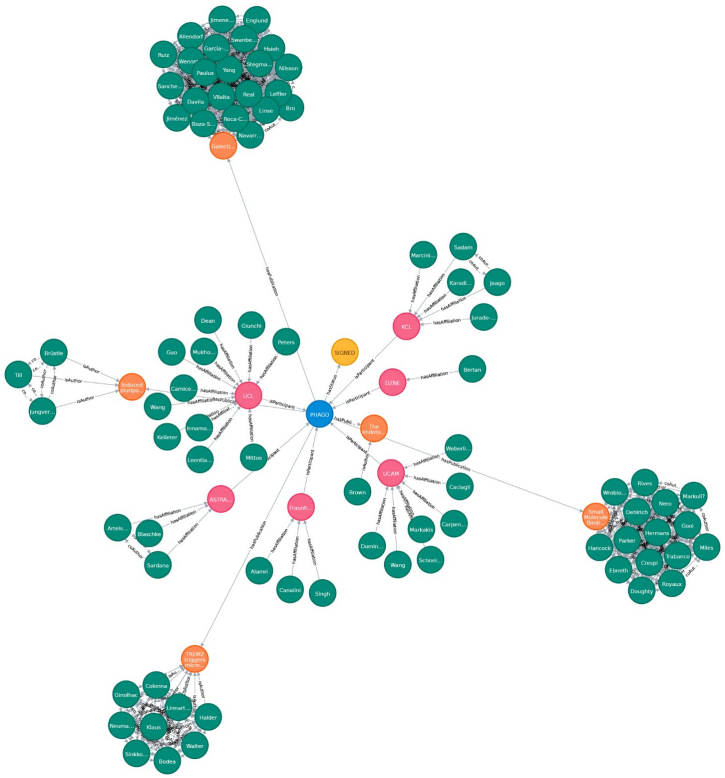


Figure 8: This illustration shows several social networks (left). They may contain actors, organizations, institutes, clubs and other actor-based entities. These actors are linked with entities and annotations from other formal structures, for example locations, attributes or – in case of literary studies – entities of characterizations, point of view, story plots, etc. They are the link to quantitative data, for example measures (age) or even survey or clinical data.

Example 6.1

We will start with a very generic example, see Figure 8. On the left, this illustration shows several social networks. They may contain actors, organizations, institutes, clubs and other actor-based entities. They share both internal as well as external relations (e.g. a person might belong to a club). These actors are linked with entities and annotations from other formal structures, for example locations and attributes. They form another context layer, whereas the actors form a context for these entities. In case of literary studies it is possible to add entities for characterizations, point of view, story plots, etc. These formal structures are the link to quantitative data, for example measures (age) or even survey or clinical data.



Example 6.2

Another example was presented by Dörpinghaus and Jacobs (2020). Here, a social network of scientists was implicitly generated with data from other sources. The study used bibliographic data from DBLP and PubMed which lead to a document layer, an author layer, a journal or venue layer etc. Limited to this data, the analysis would be limited to the study of co-authorship. By adding a project layer, we get additional information about institutes and companies, affiliated actors and other publications.
It is easy to see that we can now answer more questions on different layers. For example: Which people have collaborated in projects but have not published together or have published different findings? Which people might be a good choice for collaborating on a particular topic?

Example 6.3

A third example are webbased learning platforms. It gets more and more common to use online-learning-platforms and the amount of reliable data for its analysis is growing. For platforms using different types of metadata and user-log data different methods of learning analytics and gnowledge graphs can be used to prove educational assumptions on online learning and learning in digital environments.

Definition 6.4 (Social Network)

A *Social Network* is a Graph $G = (V, E)$ with vertices (nodes) $v \in V$ and edges (relations) $e \in E$. Both edges and vertices are part of previously well-defined categories, $V \subseteq C_1 \cup C_2 \cup ... \cup C_n$ and $E \subseteq R_1 \cup R_2 \cup ... \cup R_m$ The network G may either be directed or undirected.

Example 6.5

For example we may have 3 categories of nodes C_i containing actors, organizations and location: $V \subseteq C_1 \cup C_2 \cup C_3$.

> *[T]he problem of incomplete sources bedevils any historical network research, regardless of which genre of sources s/he bases his research on. (Rollinger 2021, 25)*

Thus, every node v and edge e in G has a source or origin which we may denote with $src(v)$ or $src(e)$ with

$$src : V \cup E \to \mathbb{S}$$

Here, $\mathbb{S}$ denotes the discrete space of sources.

We may also need to add a measure for uncertainty. For every node v and edge e in G we may denote this with $q(v)$ or $q(e)$ with

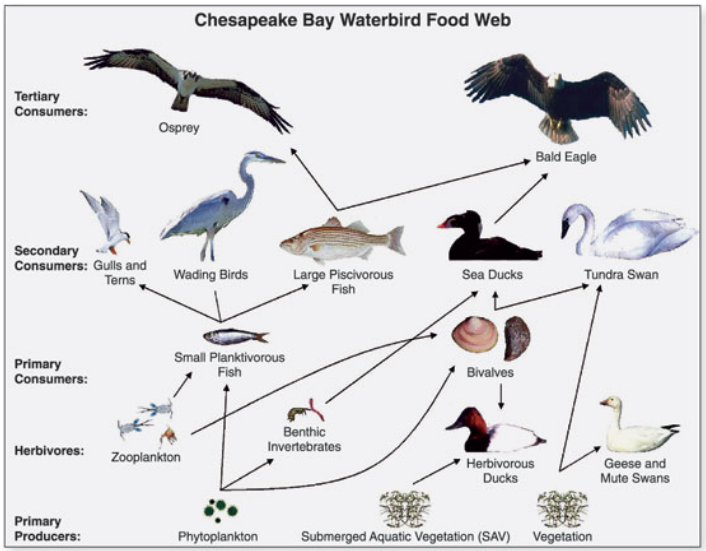
$$q : V \cup E \to (0, 1)$$

Here, 0 labels a very uncertain, 1 a very certain node or edge. This may be used to label hypothetical findings or it may be used for experimental analysis.

- Not all edge labels are necessary in every use case (SNA, HNA, RNA, Twitter analysis, ...).
- We will come back to them when discussing strong and weak ties, homophily and network dynamics.
- After that, we will discuss random graphs and in particular network types and their properties.

- Social Network Analysis is used for better understanding of society, social movements, personal networks and even to study epidemics.
- Another example: Where Social Networks rely on the species human beings, an *Ecological Network* describes all biotic interaction in a distinct ecosystem.
- Because nodes are not connected with individuals but with species, it can be seen as a deeper model of food chains.

  The main question is: how dependent are ecosystem fragility and persistence (two types of stability; Pimm 1991) on graph architecture? [...] As we will see, these networks display the robustness expected of long-tailed distributions of connections, but also a high fragility against selective species removals in terms of, first, food-web fragmentation into disconnected sub-webs, and second, secondary extinctions (i.e. species that become extinct due to the removal of other species. (Sole 2001: 406)



- We can now compute different properties of the network G.
- For example the *complexity* $c(G)$ of a network G is the average number of edges between nodes, indicating the link per species. Thus we have

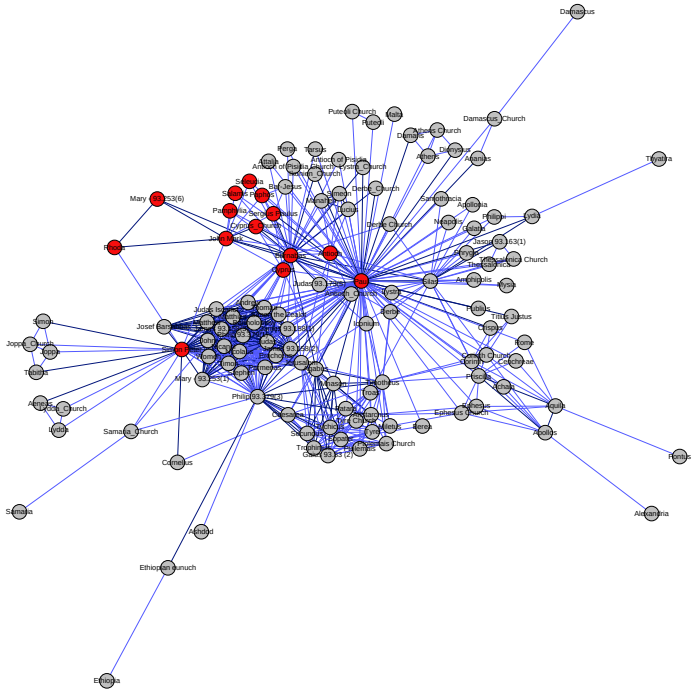
$$c(G) = \frac{1}{|V|} \sum_{v \in V} deg(v).$$

- The complexity depends on the taxon richness and topology and dynamics of interactions between the species.

- The *connectance* $C(G)$ of an Ecological Network N is defined as

$$C(G) = \frac{|E|}{|V|^2}.$$

- Observe that the maximum number of edges or links is the square of the number of nodes $|V|^2$, and that the connectance should not be confused with the graph connectivity which has a completely different meaning.
- By the above definition, C is thus in the range $[0, 1]$ and usually far from 1 since Ecological Networks are usually sparse.



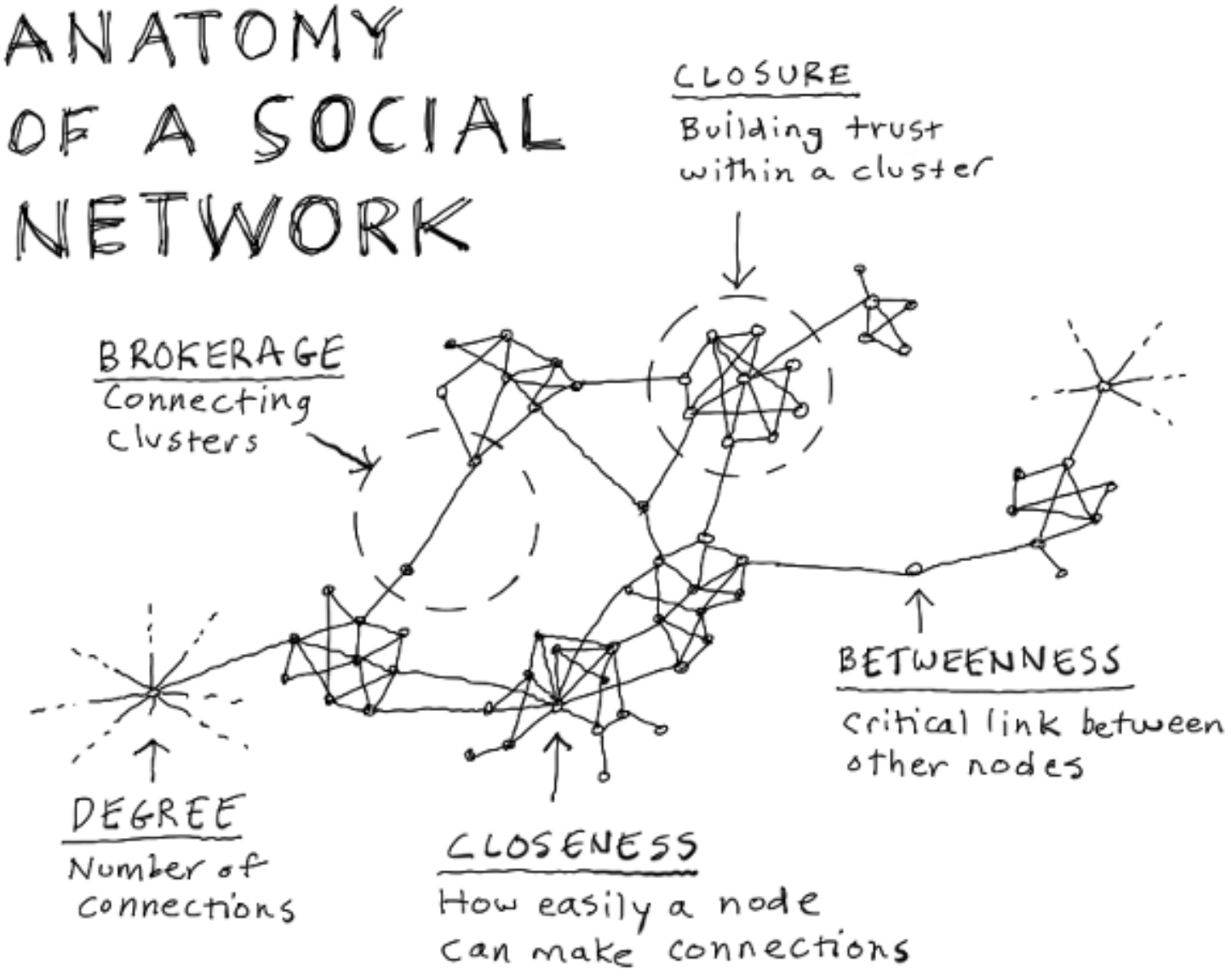
- The *neighborhood* gives information about the connectedness of an actor in the network.
- This can be useful to illustrate the direct influence of an actor within the complete network, especially for actors with a high node degree.
- But it is obvious that the amount of relations does not necessarily give a good idea on their quality or how we could use these relations.

- While the node degree is often used as a measure to create random graphs, it is in general not a good measure in order to analyze particular actors in networks, see (Jackson, 2010).
- Nevertheless, the *degree centrality* for a node $v \in V$ is given by

$$dc(v) = \frac{deg(v)}{n-1}$$

- The output value ranges between 0 and 1 and gives a reference to the direct connections.
- As discussed, it omits all indirect relations and in particular the node's position in the network.

- We will now discuss one more property to evaluate nodes and their position in the networks.
- These properties can be used to calculate statistical parameters, so-called *centrality measures*.
- They answer the question “Which nodes in this network are particularly significant or important?”.



Gray & Vander Wal (2014)

- *Betweenness* analyzes critical connections between nodes and thus gives an indication of individuals that can change the flow of information in a network. This measure is based on paths in a network:
  Much of the interest in networked relationships comes from the fact that individual nodes benefit (or suffer) from indirect relationships. Friends might provide access to favors from their friends, and information might spread through the links of a network. (Jackson, 2010)

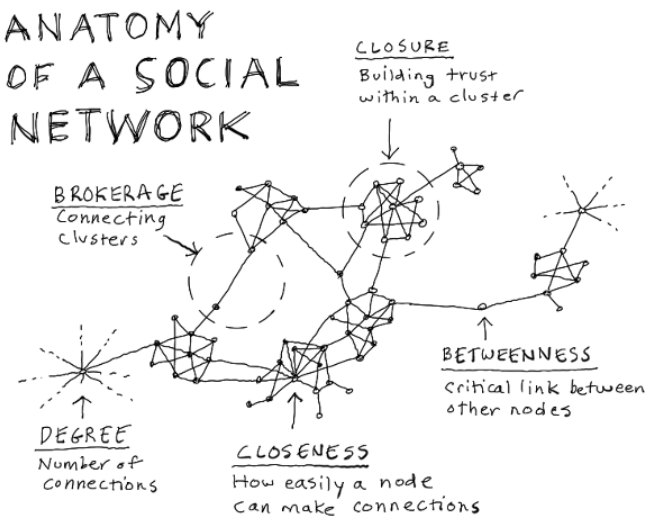
Definition 6.6 (Betweenness centrality (Freeman 1977, White 1994))

Let $G = (V, E)$ be a graph with a node $v \in V$. Then the betweenness centrality is given by

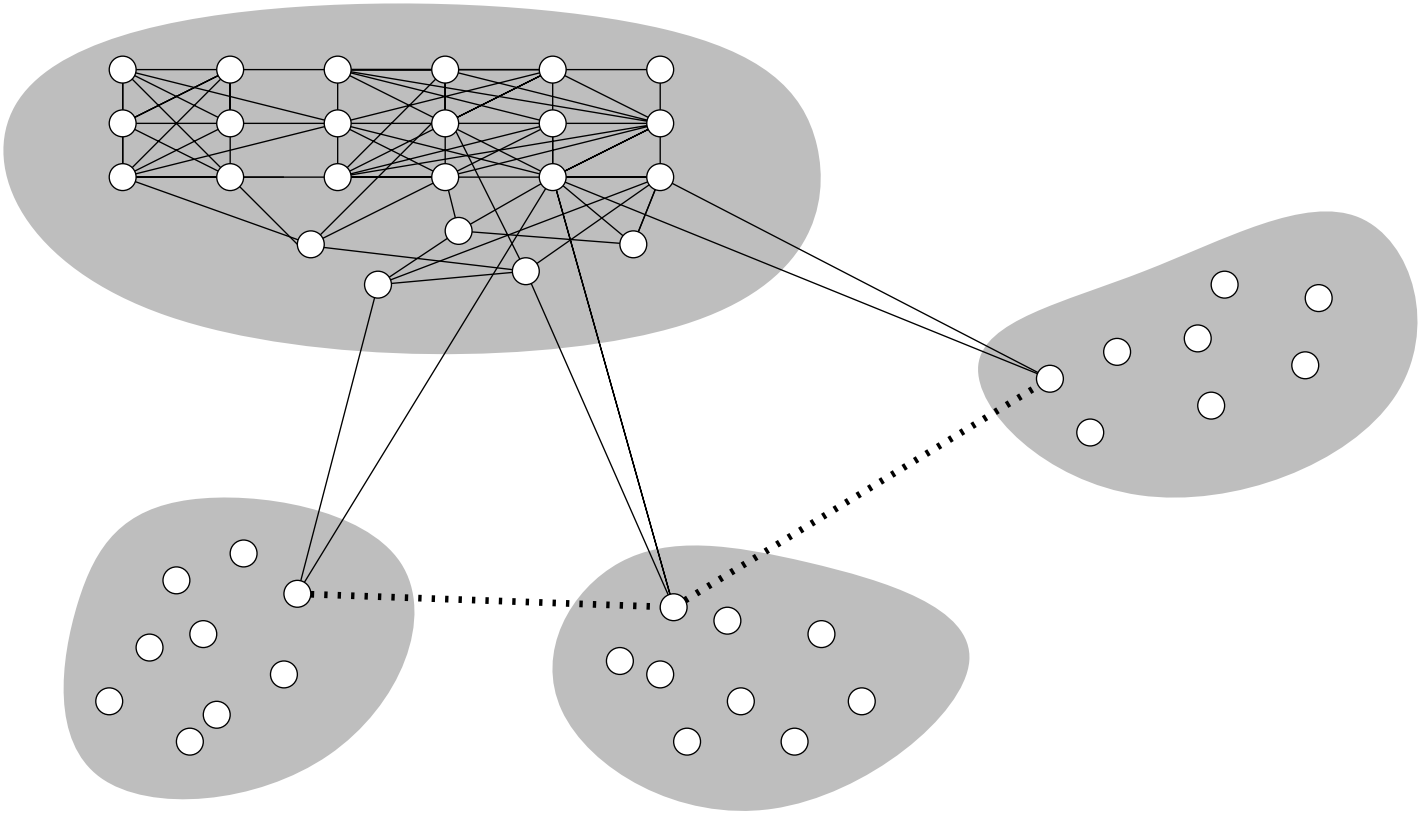
$$bc(v) = \sum_{k \neq j, v \neq k, v \neq j} \frac{P_v(k,j)}{P(k,j)} \cdot \frac{2}{(n-1)(n-2)}.$$

- Given a node v, it calculates the number $P_v(k,j)$, that is, the number of all shortest paths in a network for all beginning and ending nodes $k, j \in V$ that pass through v.
- If $P(k,j)$ denotes the total number of paths between k and j, the importance of v is given by the ratio of both values.

- This parameter allows an analysis of the critical links and how often a node lies on such a path.
- This centrality measure thus answers the questions whether a node can change the flow of information in a network or whether it is a bridge between other nodes.
- While betweenness assumes network flows to be like packages flowing from a starting point to a destination, other measures consider multiple paths: For example, the so-called *eigenvector centrality*.
- Less popular measures are Katz prestige, and Bonacich's measure. It has been shown that these measures are closely related, see (Ditsworth 2019).



- Remember: a subgraph $C \subset G$ is called a *clique* if all nodes in C are pairwise connected.
- Thus, a clique refers to the colloquial use of this word, but in real-world networks there are only very limited cliques where all nodes are connected.
- Scholars are more interested in dense connected subnetworks which are called communities. We will discuss these structures in detail in the next section.



- A stable set refers to a *structural hole* introduced by Burt (2004).
- From a mathematical perspective, a clear definition is missing and most scholars are following Burt's vague description.
- Burt states, "that people who stand near the holes in a social structure are at higher risk of having good ideas."(:349)
- They are defined by non-connected areas (a stable set is "the hole") and people bridging them (the brokerage).
- But the impact of this structure is discussed extensively in literature, see for example Cowan (2007) or the discussion if "actors with closed networks [...] are disadvantaged in terms of information and control benefits"(Kilduff 2010: 329).
- Beside these general questions it is also a yet unresolved questions weather an existing network structure describing a particular characteristic also implies that it is the reality.

We will now consider two graph structures to take a closer look at their impact on the error measures.

- We will now evaluate how graph structures and in particular measures change when additional information are stored in extra layers.
- We partition a graph into an uncolored part that contains the 'original' data and into a part with blue nodes in which novel 'extra' data stored.
- These blue nodes simulate one or more new layers in the knowledge graph.
- Thus, given a random graph $G = (V, E)$, a next step comprises a probability p_b for blue nodes which leads to a graph G with blue nodes $B \subset V$.
- We first compute the centrality measures for all nodes in $V \setminus B$ in the graph $G = (V, E)$.
- Then we compute those measures for all nodes in $G \setminus B$, this time in the Graph $G \setminus B = (V \setminus B, E)$.

- Let $dc(G)$ be the vector containing the degree centrality measures for all nodes v in G in descending order, where - after the computation of $dc(v)$ for all $v_1, \ldots, v_n \in V(G)$ - the values for all $v \in B$, that is, v_i with $i = n - |B| + 1, \ldots, n$, are deleted.
- Hence,

$$dc(G) = (dc(v_1), dc(v_2), ..., dc(v_{n-|B|}))$$

  with $dc(v_j) \geq dc(v_{j+1})$ for all $j \in \{1, \ldots, n - |B|\}$.
- We may do the same for bc.

- While comparing two vectors, we are interested in two values.
- The first one is the total number of misordered elements, that is, the total number of positions on which the elements differ from each other.
- The second value that we compute in order to compare two vectors is the number of moved elements.
- For this we count those elements that have a different predecessor and / or successor in the first vector compared to the second one.

Example 6.7

Let $c_1 = [1, 2, 3, 4, 5]$, $c_2 = [5, 3, 2, 1, 4]$ and $c_3 = [1, 5, 2, 3, 4]$. If c_1 is the original ordering, we see that c_2 has a totally different order. In c_3 the entry 5 is moved, but the rest of the list is unchanged, although still 4 elements are on the wrong location. Hence, the number of misordered elements in c_1 compared to c_2 is 5. The number of moved elements is 5 and 1.

- To identify both errors, we first define function e:

$$e(i, c_1, c_2) = \begin{cases} 0 & c_1^i = c_2^i \\ 1 & c_1^i \neq c_2^i \end{cases}$$

- Let x be an element contained in every c_u, $u \in \mathbb{N}$.
- Then $p(x, c_u)$ denotes the predecessor of element x in c_u and $s(x, c_u)$ denotes the successor of x in c_u.
- If x is the first element in c_u, then $p(x, c_u) = \emptyset$.
- If x is the last element of c_u, then $s(x, c_u) = \emptyset$.
- With these definitions, we define e_N:

$$e_N(x, c_1, c_2) = \begin{cases} 1 & \text{if } p(x, c_1) = \emptyset \text{ and } s(x, c_1) \neq s(x, c_2), \\ & \text{or } s(x, c_1) = \emptyset \text{ and } p(x, c_1) \neq p(x, c_2), \\ & \text{or } s(x, c_1) \neq s(x, c_2) \text{ and } p(x, c_1) \neq p(x, c_2), \\ 1/2 & \text{if } s(x, c_1) \neq s(x, c_2) \text{ and } p(x, c_1) = p(x, c_2), \\ & \text{or } s(x, c_1) = s(x, c_2) \text{ and } p(x, c_1) \neq p(x, c_2), \\ 0 & \textit{otherwise}. \end{cases}$$

With this, we define two error measures ϵ and ϵ_N:

$$\epsilon(c_1, c_2) = \sum_{i=1}^{n} e(i, c_1, c_2)$$

$$\epsilon_N(c_1, c_2) = \sum_{x \in c_1} e_N(x, c_1, c_2)$$

Example 6.8

Let's reconsider the previous example: Recall that $c_1 = [1, 2, 3, 4, 5]$, $c_2 = [5, 3, 2, 1, 4]$ and $c_3 = [1, 5, 2, 3, 4]$. Then, $\epsilon(c_1, c_2) = 5$ and $\epsilon_N(c_1, c_2) = 5$. Moreover, $\epsilon(c_1, c_3) = 4$ and $\epsilon_N(c_1, c_3) = 2.5$.

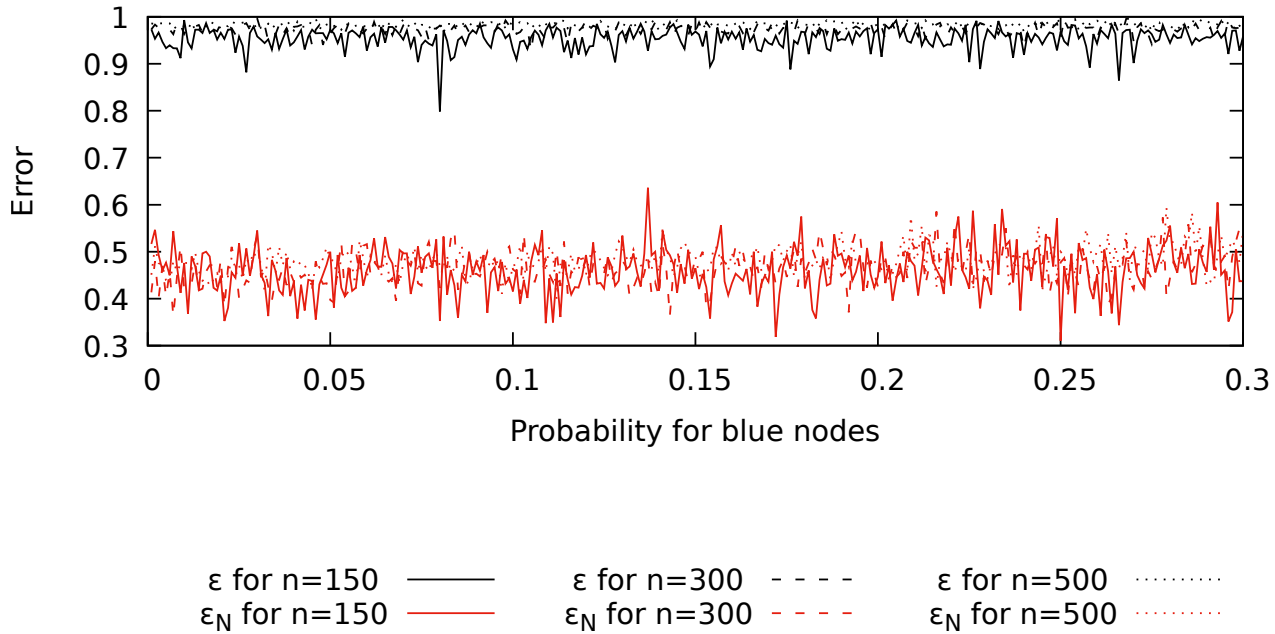


Figure 9: Degree Centrality errors for scale-free random graphs ($n = 150$, $n = 300$ and $n = 500$) for different values of p_B between 0 and 0.3.

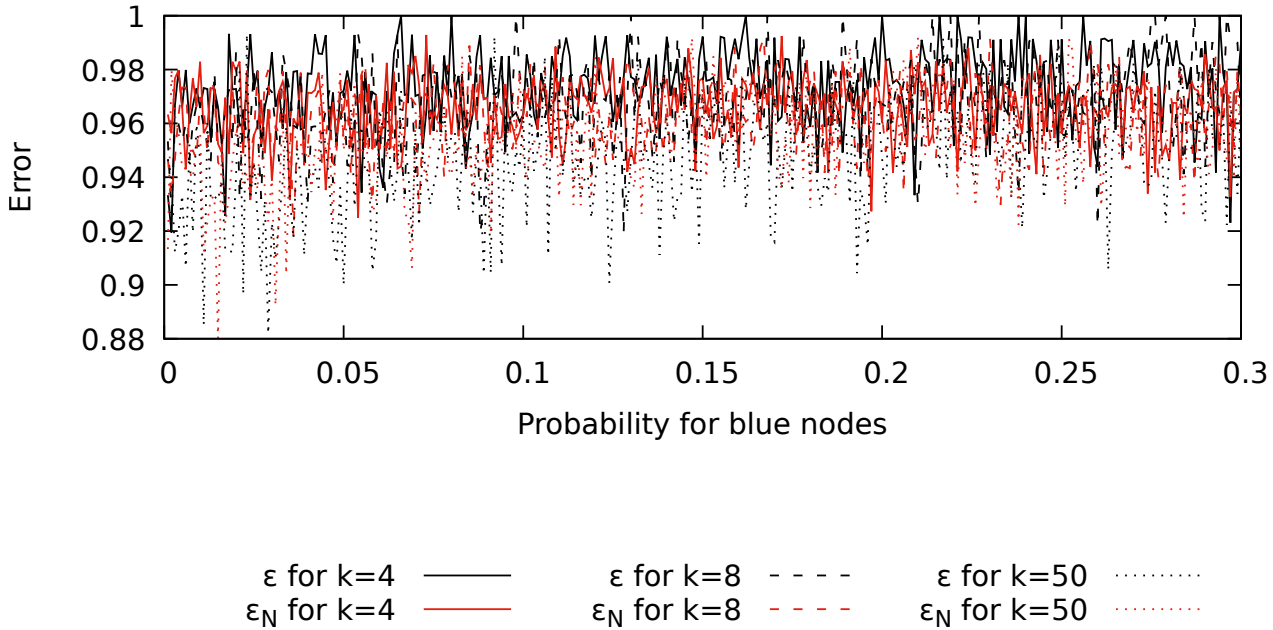


Figure 10: Degree Centrality errors for Newman-Watts-Strogatz small-world random graph ($n = 150$, $k \in \{4, 8, 50\}$) for different values of p_B between 0 and 0.3.

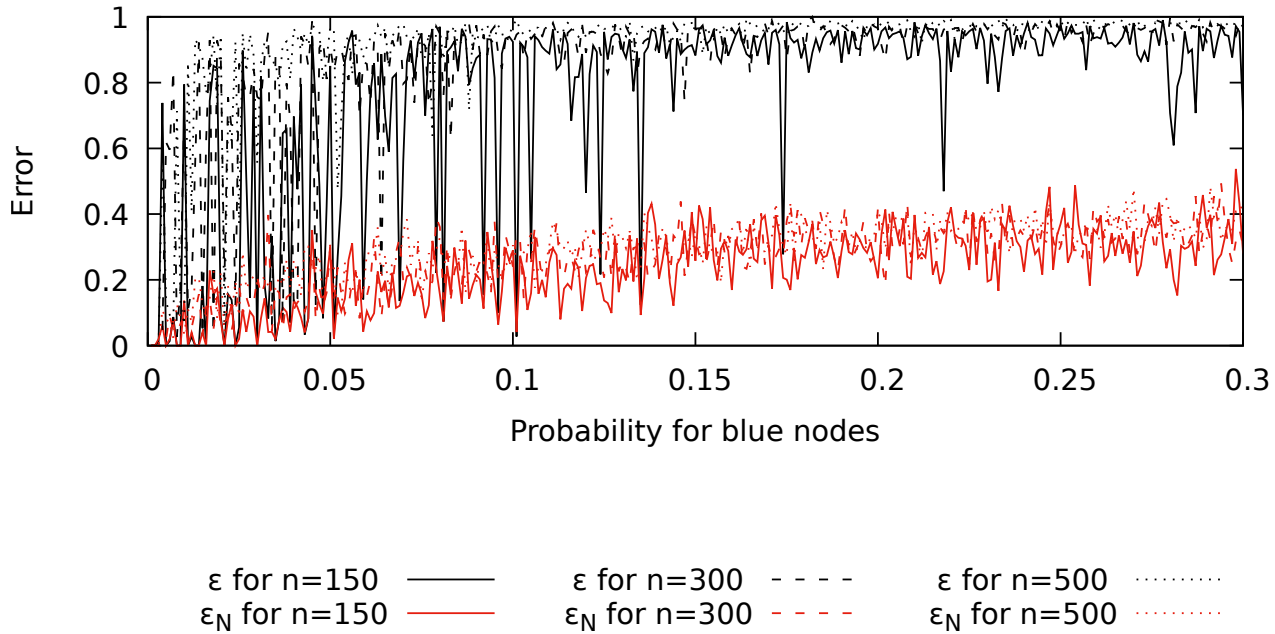


Figure 11: Betweenness Centrality errors for scale-free random graphs ($n = 150$, $n = 300$ and $n = 500$) for different values of p_B between 0 and 0.3.

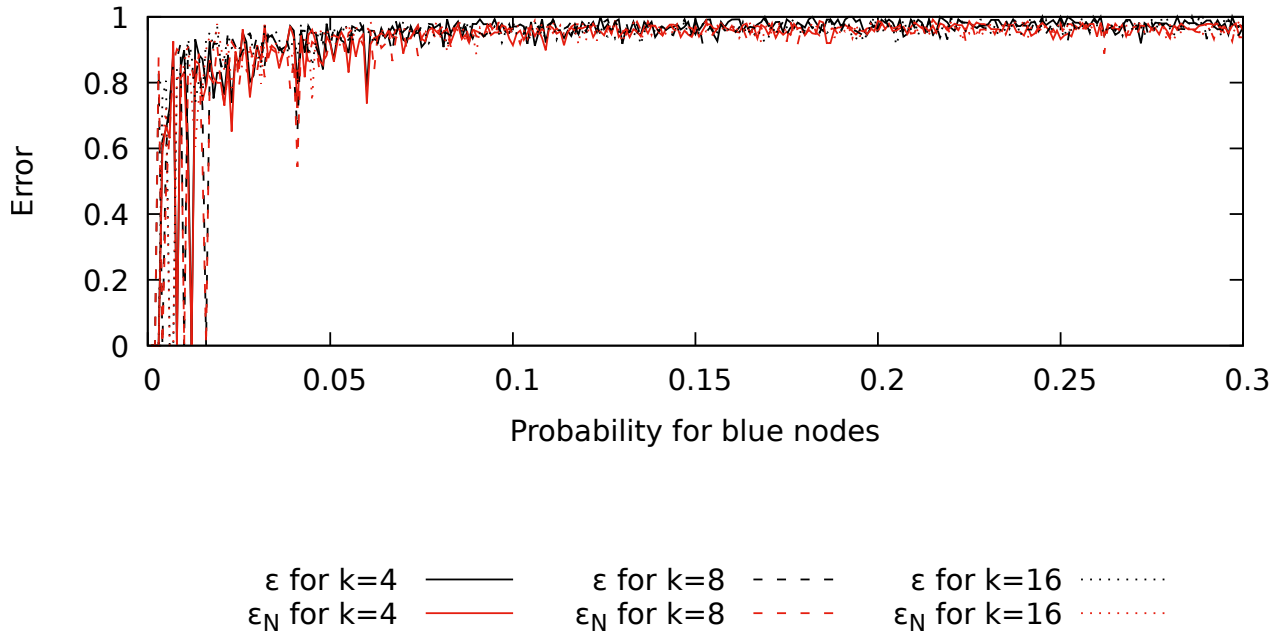


Figure 12: Betweenness Centrality errors for Newman-Watts-Strogatz small-world random graph ($n = 150$, $k \in \{4, 8, 50\}$) for different values of p_B between 0 and 0.3.

- Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$.
- The nodes in $G \setminus B = (V \setminus B, E)$ are denoted by $v_1, \ldots, v_{n-|B|}$ while the nodes in B are denoted by $v_{n-|B|+1}, \ldots, v_n$.
- We further assume that $G \setminus B$ is still connected.

Lemma 6.9

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let C_k be a clique in G with k nodes and let $u \in C_k$. Then

$$\epsilon(dc(G), dc(G \setminus B)) \leq n - 1 - \min_{v \in N(u)} p_{dc}(v)$$

holds.

Proof.

Let $a_1 = dc(G)$ and $a_2 = dc(G \setminus B)$. The only nodes which are affected by a decreasing degree centrality are those in the neighborhood $N(u)$ of the blue node u, since for $v \in N(u)$, only one node in the neighborhood of v is removed in $G \setminus B$ compared to G. Thus,

$$a_1^{p(v)} = a_2^{p(v)} - 1 \quad \forall v \in N(u)$$

holds. Observe that $\min_{v \in N(u)} p_{dc}(v)$ denotes the smallest position in $dc(G)$ of a node in $N(u)$ (that is, the highest ranked neighbor of u in $dc(G)$). All nodes in $dc(G)$, that are higher ranked are not affected by the deletion of u. Recall that $dc(G)$ only has $n - |B| = n - 1$ entries. Thus, at most $n - 1 - \min_{v \in N(u)} p_{dc}(v)$ nodes change their position in $dc(G \setminus B)$ compared to $dc(G)$. □

Lemma 6.10

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let C_k be a clique in G with k nodes and let $u \in C_k$. Then

$$\epsilon_N(dc(G), dc(G \setminus B)) \leq k - 1$$

holds.

Proof.

Let $a_1 = dc(G)$ and $a_2 = dc(G \setminus B)$. Again, the only nodes which are affected by a decreasing degree centrality are those in the neighborhood of the blue node, that is the set $v \in N(u)$. Here, only one node in the neighborhood of these nodes is removed in $G \setminus B$. While the internal order of all nodes in $G \setminus \{C_k \setminus \{u\}\}$ does not change and the internal order of the $k-1$ nodes in $C_k \setminus \{u\}$ remains untouched as well, at most the $k-1$ nodes in $C_k \setminus \{u\}$ are shifted to a certain degree to the right, since their value in $dc(G \setminus B)$ decreased compared to $dc(G)$. Every vertex in $C_k \setminus \{u\}$ hence contributes at most 1 to the sum computed in $\epsilon_N(dc(G), dc(G \setminus B))$, which leads to the upper bound $k-1$. □

Since betweenness centrality is also affected by the global structure of the graph, counting all shortest paths, the situation is slightly different.

Lemma 6.11

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let C_k be a clique in G with k nodes and let $u \in C_k$. Then

$$\epsilon(bc(G), bc(G \setminus B)) \leq \begin{cases} 0 & \text{if } d(u) = k - 1 \\ \sum_{w \neq y} P_u(w, y) & \text{otherwise.} \end{cases}$$

We now consider the special case in which the only existing blue node has degree 1.

Lemma 6.12

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $N(u) = \{v\}$. Then

$$\epsilon\left(dc\left(G\right), dc\left(G \setminus B\right)\right) \leq p_{dc}(v)$$

holds.

Proof.

Exercise. □

Lemma 6.13

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $|N(u)| = 1$, that is, $d(u) = 1$. Then

$$\epsilon_N(dc(G), dc(G \setminus B)) \leq 2$$

holds.

Lemma 6.14

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $N(u) = \{v\}$ and let

$$t = \max\left\{n - 1, \sum_{w \neq y} P_v(w, y)\right\}.$$

Then

$$\epsilon(bc(G), bc(G \setminus B)) \leq t$$

holds.

Lemma 6.15

Let $G = (V, E)$ be a graph with $|V| = n$ and blue nodes $B \subset V$ with $B = \{u\}$ where $G \setminus B$ is still connected. Let further $N(u) = \{v\}$ and let
Let

$$z = \max\left\{2\sum_{w \neq y} P_u(w, y), 2\sum_{w \neq y} P_v(w, y)\right\}.$$

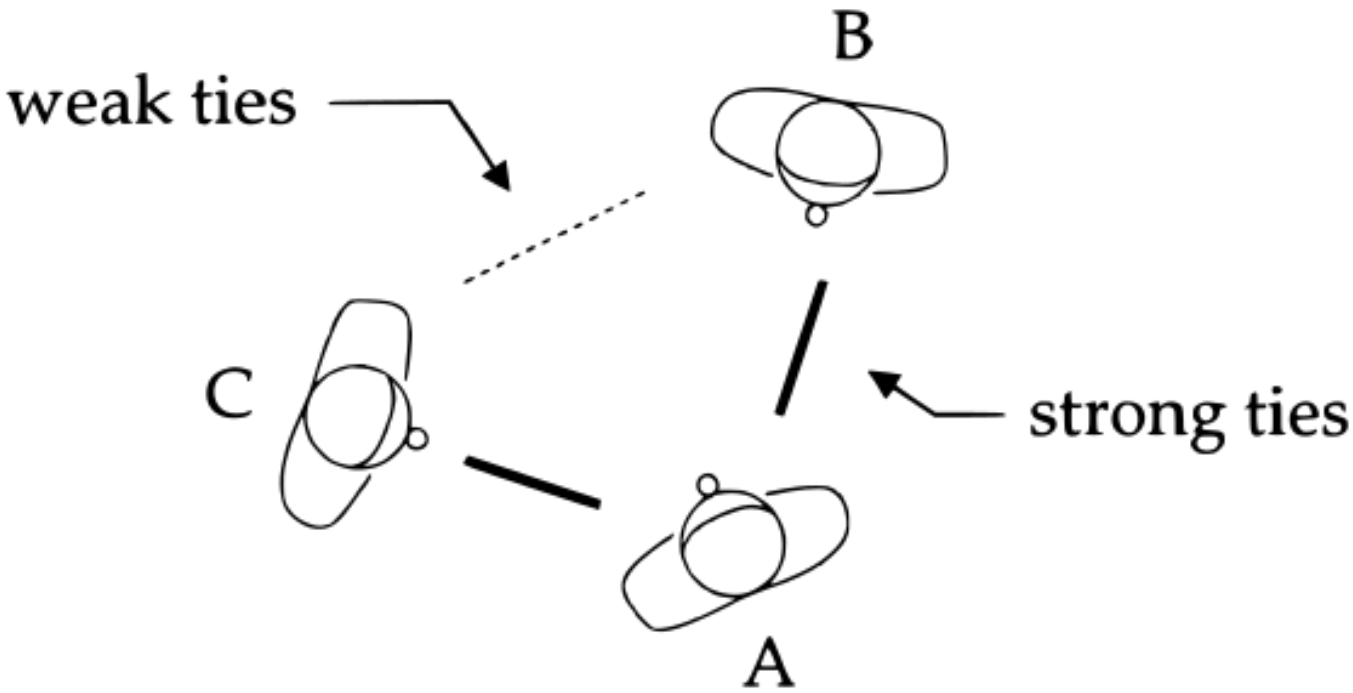
Then

$$\epsilon_N(bc(G), bc(G \setminus B)) \leq z$$

holds.

- These error estimations are not sharp.
- In addition, if the size of B increases, it will be even more challenging to specify the error rates.
- We could show that blue nodes in clusters have a great influence on both ϵ and ϵ_N while those nodes with a small neighborhood have a rather small influence on ϵ_N.
- This gives a first idea why in general scale-free networks are more robust regarding ϵ_N.
- The degree centrality is only influenced by local structures but in general the errors are higher while the betweenness centrality is in general more complex and the results of this paper can only give some hints, but further research needs to be done here.

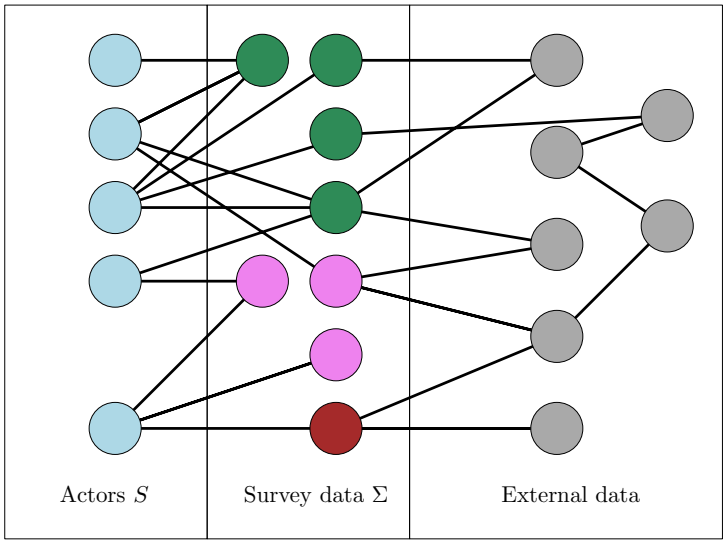
- A connection between two persons can either be strong or weak. This theory goes back to the american sociologist Mark Granovetter.
- In the 1970s, he claimed that information flow, politics, marketing and so on goes mainly through weak ties.
- Thus, not the connectance of an actor (which means the number of persons connected to) seems to be important, but the number of weak ties and connections through those ties a person has.
- This approach is under high research and highly discussed.



- Granovetter's categorization of ties is also highly relevant to the transitivity of ties in triangular form. Here, for three individuals A, B, and C, it holds that the stronger the friendship relationship between individuals A and B and individuals A and C, the more likely that B and C know or will know each other.
- This is denoted as 'Triadic Closure'.
- One way to analyze social networks is to examine the density of the respective network.
- In a *dense and multiplex network*, everyone knows everyone else;
- in a *thin and so-called uniplex network*, not all people know each other and there is usually only one type of relationship between people.
- However, only those sub-networks that are particularly dense and multiplex would be considered significant.
- Therefore, especially non-existing or weak relationships were investigated.

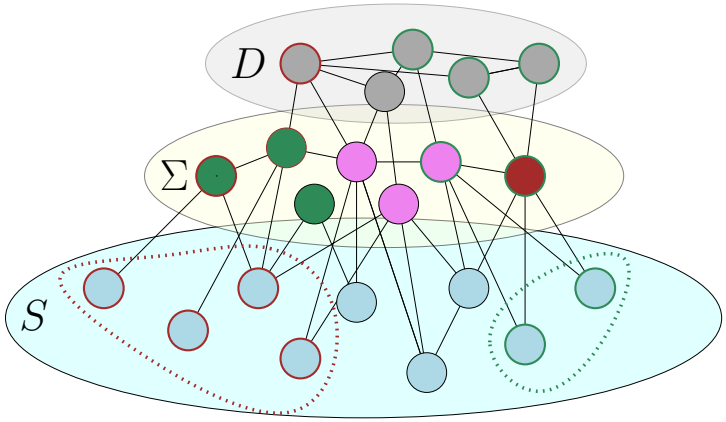
- Whether weak or strong ties are important for the spread of new ideas is also heavily discussed.
- Two Examples:
- Collar, finds it much more important to observe that "*everybody* is both a weak and a strong tie, that identification as such depends on perspective, and that these classifications are flexible and subject to change."(Collar 2014: 36). She calls for a particular focus on strong ties.
- She is contradicted by Schweizer (1996: 118f), who sees weak ties as bridges between different subnetworks.

- Another yet open question is how to distinguish between strong and weak ties.
- Besides, we may also consider negative relations (cf. Everett 2014: 111).
- It may be a good solution to assume either *no*, a *negative* or a *weak* or *strong* relations.
- However, although these suggestions may make sense form a sociological perspective, they clearly show the interdisciplinary challenges. In particular, they show that graph theory can only help to solve problems which have been well-defined beforehand.



- Lee (2019) presented a study on the perception about the frequency of attributes and measures in the ego-network of actors.
- They claim, that the perception biases are related solely to the structure of social networks, in particular “on the level of homophily and its asymmetric nature, as well as on the size of minority group”(:1).
- First, a social network model was generated where the homophily parameter (regulating the probability of connections between actors from minority and majority groups) and minority-group size was adjusted according to their theoretical model.
- When balancing the aggregation of perception within an ego-network this study claims that perception biases can be reduced in heterophilic, but not in homophilic networks.

- In this case, the Knowledge Graph $G = (V, E)$ comprises a social network layer $S = (V_S, E_S)$ and survey data $\Sigma = (V_\Sigma, E_\Sigma)$.
- Beside of that we may also add more data layers, for example explaining survey data or adding more quantitative or qualitative data to S.
- We may now assume that one ore more layers are partly incomplete, for example that S is missing social relations.
- There are several questions: Can we predict relations in S? Is it possible to apply methods on S without the missing data, for example community detection or computing centrality measures?



- It is quite obvious that the model introduced by Lee (2019) could be used to re-predict relations in S.
- But there is a serious shortcoming:
- The model was build on real-world scenarios to explain the perception biases within an ego-network.
- But unfortunately the survey data itself does only refer back to a single actor network S, not to a complete ego-network or a network.
- We may denote the hidden, complete social network with $\hat{S}$.
- Thus, this approach would only be feasible once we have survey data for such a complete social (sub-)network.

- On the other hand, it is possible to apply methods like Community Detection on G.
- But given one particular method, the result on G, S and $\hat{S}$ may vary.
- Applying the method on S wouldn't be useful, since the relations (and actors not participating on the survey) are missing.
- But assuming the complete (sub-)network $\hat{S}$, we would expect that the results on $G|_S$ are the same or comparable to those on $\hat{S}|_S$.

- In general a *community* is a subset $C \subseteq V$ of nodes.
- The corresponding graph-theoretical problem is called *graph partition*. Other network researches use a different nomenclature: Easley (2019: 70) for example name communities 'regions'.

Definition 6.16

A *community structure* is a set of communities $C_1, ..., C_n$ so that $C_1 \cup ... \cup C_n = V$. If these communities are pairwise disjoint, we have hard borders between these sets. If nodes or actors may belong to multiple communities we may speak of soft or overlapping community detection.

- But beside of this definition we lack a detailed understanding of *what* a community is.
- Several prominent definitions exist and from a mathematical perspective they can't be merged. Thus there is a little arbitrariness. Are communities "tightly knit groups with dense connections among their members"(Newman 2006:553)
- or solely unconnected margin actors?
- Jackson (2019:449) summarizes:

  > *As we vary what a community represents, the optimal method for identifying communities will correspondingly change. In particular, it is important to have an idea of how community structure affects network formation.*

- Discussing how nodes may be playing a similar role is still an ongoing discussion in research.
- Thus, the initial research questions remains open: Are results on $G|_S$ the same or comparable to those on $\hat{S}|_S$?

- Fortunato (2010) provides a detailed overview on methods and approaches.
- We will mostly rely on those called ‘traditional methods’.
- Most research focuses on the performance of these algorithms in dense and large-scale networks.
- In general, it is unclear how these methods perform in small social networks.

- The *Girvan-Newman* algorithm was introduced in Girvan (2002).
- This approach relies on “property of community structure, in which network nodes are joined together in tightly knit groups, between which there are only looser connections.” (:7821)
- It is based on the removal of nodes with the highest betweenness.
- Thus, it does not aim at finding the strong connected centers of a network.
- The time complexity of this algorithm is very high, $\mathcal{O}((m+n)n^2)$, or $\mathcal{O}(n^3)$.
- It was shown that this algorithm does not work well for every graph class.
- In these cases, Girvan-Newmann will produce very unbalanced partitions.

Algorithm 10 Girvan-Newman

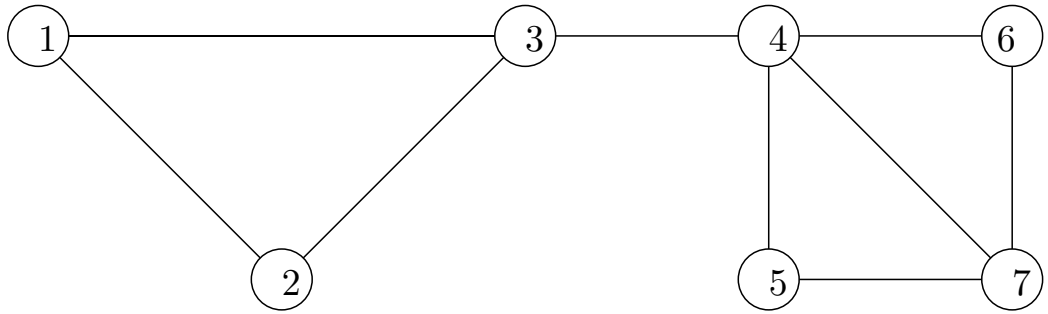
1: Computation of the centrality for all edges;
2: Removal of edge with largest centrality: in case of ties with other edges, one of them is picked at random;
3: Recalculation of centralities on the running graph;
4: Iteration of the cycle from step 2.

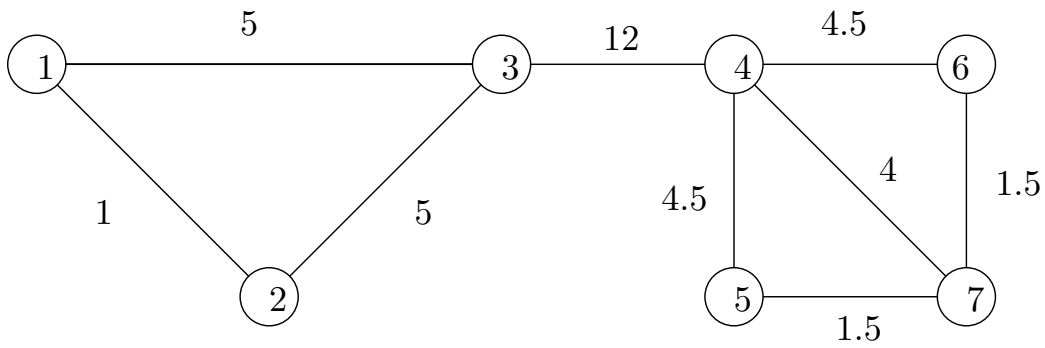
Definition 6.17

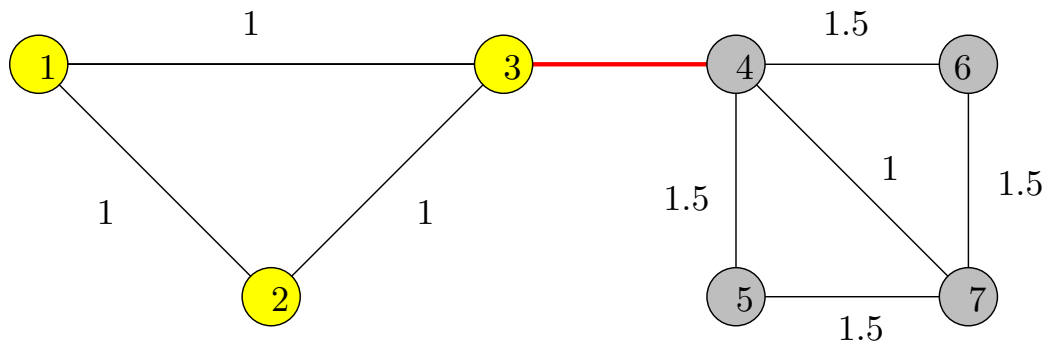
Edge betweenness is the number of shortest paths between all vertex pairs that run along the edge:

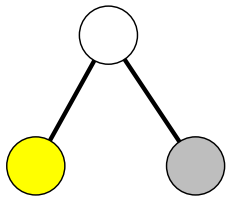
$$bc(e) = \sum_{k,j \in V} \frac{P_e(k,j)}{P(k,j)}.$$

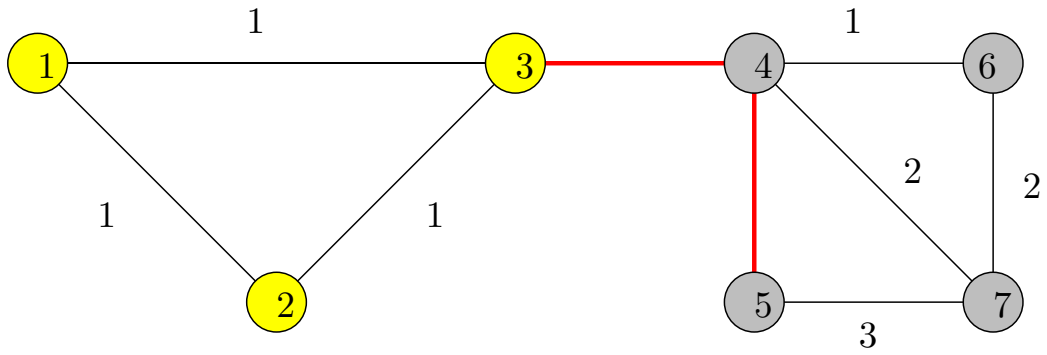
- The result of the Girvan-Newman algorithm is a dendrogram.

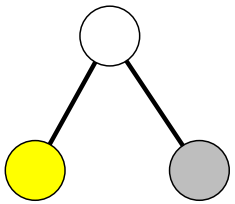


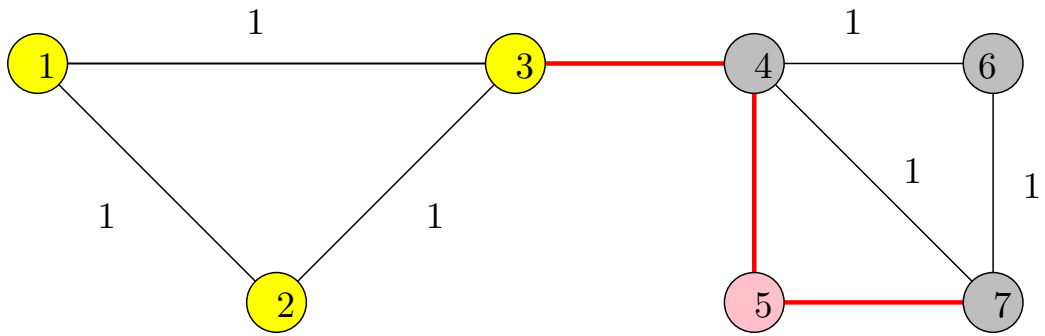

5
12
4.5
1
5
4.5
4
1.5
1.5
1
2
3
4
5
6
7

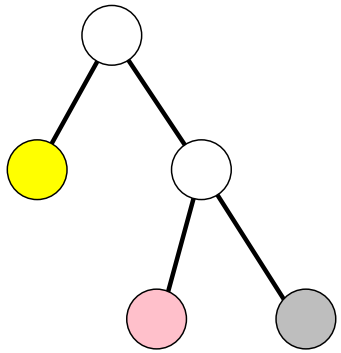

1
1.5
1
3
4
6
1
1
1.5
1
1.5
2
5
7
1.5

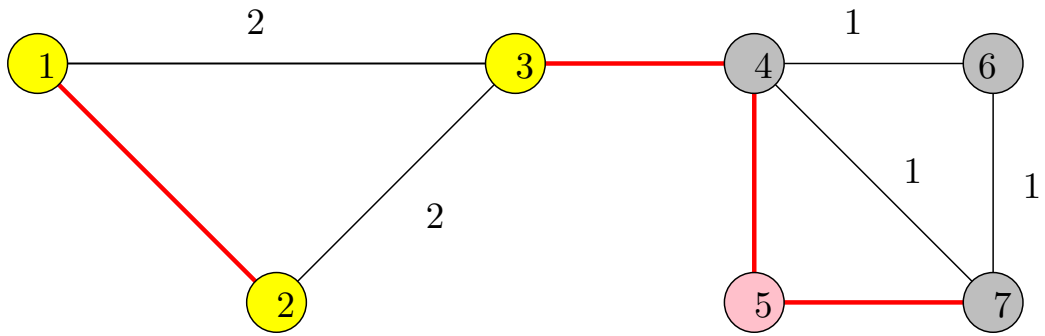


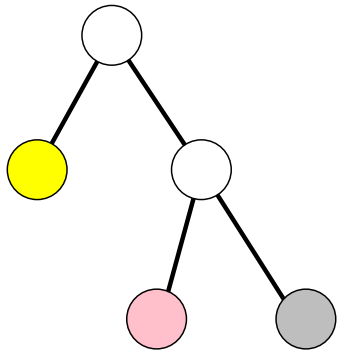

1
1
1
1
2
1
2
2
3
1
2
3
4
5
6
7

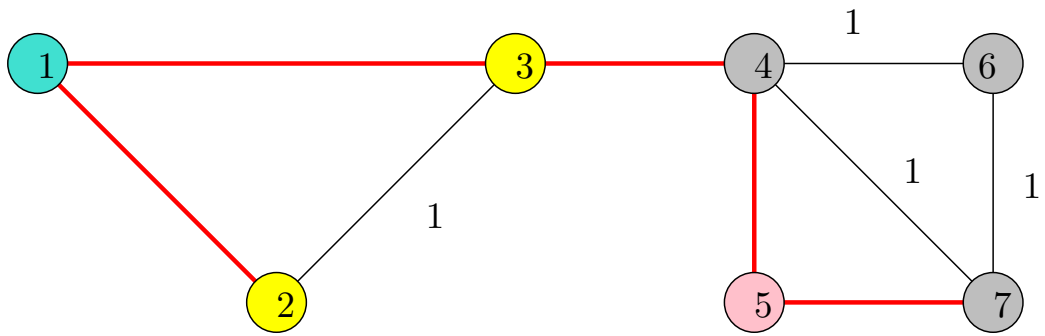


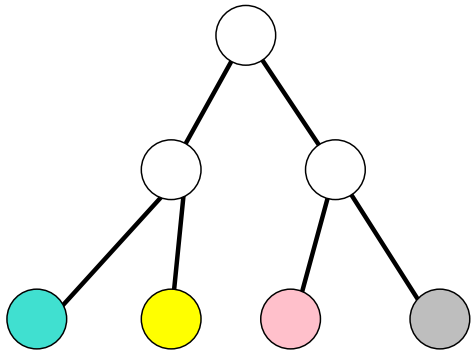

1
1
1
1
1
1
1
2
3
4
5
6
7

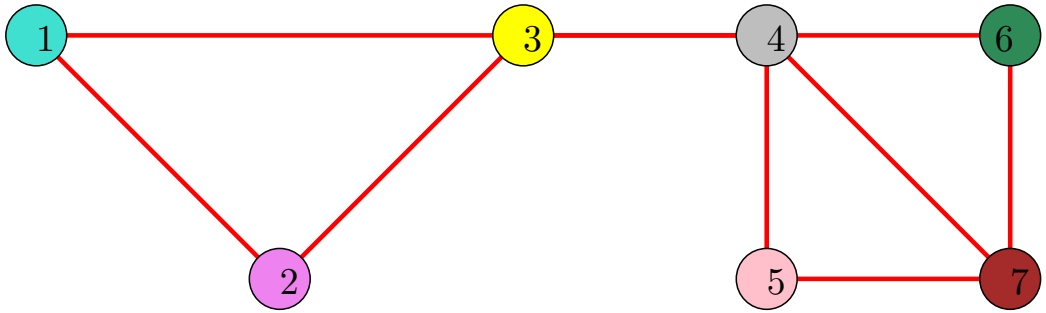


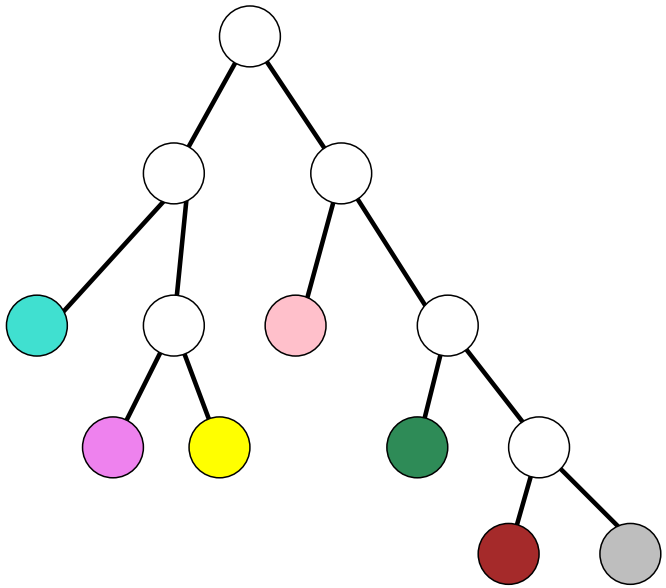

2
1
1
3
4
6
1
1
2
2
5
7




1
3
4
6
2
5
7
1
1
1
1




1
3
4
6
2
5
7

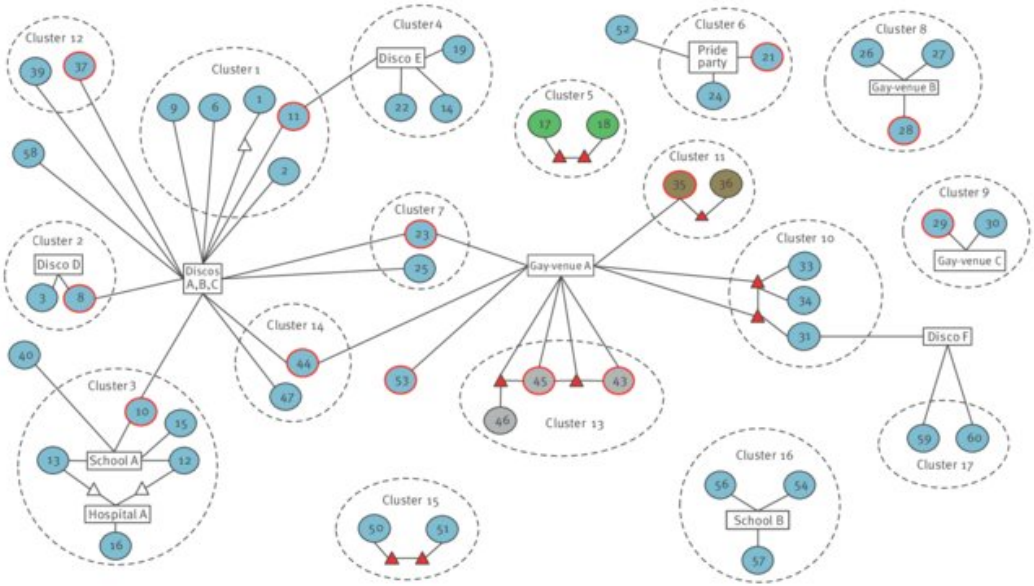


- The *Leiden* algorithm is also very popular and was first introduced by Traag (2019). as an extension of the classical Louvain algorithm.
- This approach is based on the optimization of modularity or Constant Potts Model (CPM) values in communities to the expected value.
- It is often described as Louvain approach in literature.
- Since this algorithm focuses on the inherent structure of a community in comparison to other communities it will most likely output similar communities.

- Another approach called *Fluid Communities* (FluidC) was introduced by Pares (2017).
- It is inspired by the idea of expanding fluids which will end in a stable environment and refer to communities.
- Thus, it allows to define the number of expected communities which will lead to a different number of ‘fluids’ starting in the network.
- It is a very efficient algorithm with good performance.
- This algorithm will be hard to compare with the other two approaches since we need to define a number of communities to obtain.

- Easley (2010: 78) states “it is a challenge to rigorously evaluate graph partitioning methods and to formulate ways of asserting that one is better than another – both because the goal is hard to formalize, and because different methods may be more or less effective on different kinds of networks.”
- And this is very true:
- We can’t give a proper mathematical definition of what we want to see; also due to the fact that we work exploratively and can’t provide a-priori quality measures.

- Network dynamics is a research field for the study of networks whose status changes in time.
- For example: SNA can be used to analyze epidemiological links between subgroups.



- Network Dynamics is highly related to the longitudinal modeling and analysis of data in networks.
- Thus, we will come back to this topic when talking about information networks.

- Let V be a fixed set of n elements, say $V = \{0, ..., n-1\}$.
- Our aim is to turn the set $\mathbb{G}$ of all graphs on V into a probability space, and then to consider the kind of questions typically asked about random objects:
  - What is the probability that a graph $G \in \mathbb{G}$ has this or that property?
  - What is the expected value of a given invariant on G, e.g. chromatic number?
  - What are typical structures of a given phenomenon (e.g. social networks) and how can we generate graphs with this property?

- For every potential edge $e \in [V]^2$ we define a probability space

$$\Omega_e := \{0_e, 1_e\} \text{ with } \mathbb{P}_e(\{1_e\}) := p,\ \mathbb{P}_e(\{0_e\}) := q$$

- As our desired probability space $\mathcal{G} = \mathcal{G}(n,p)$ we then take the product space

$$\Omega := \prod_{e \in [V]^2} \Omega_e$$

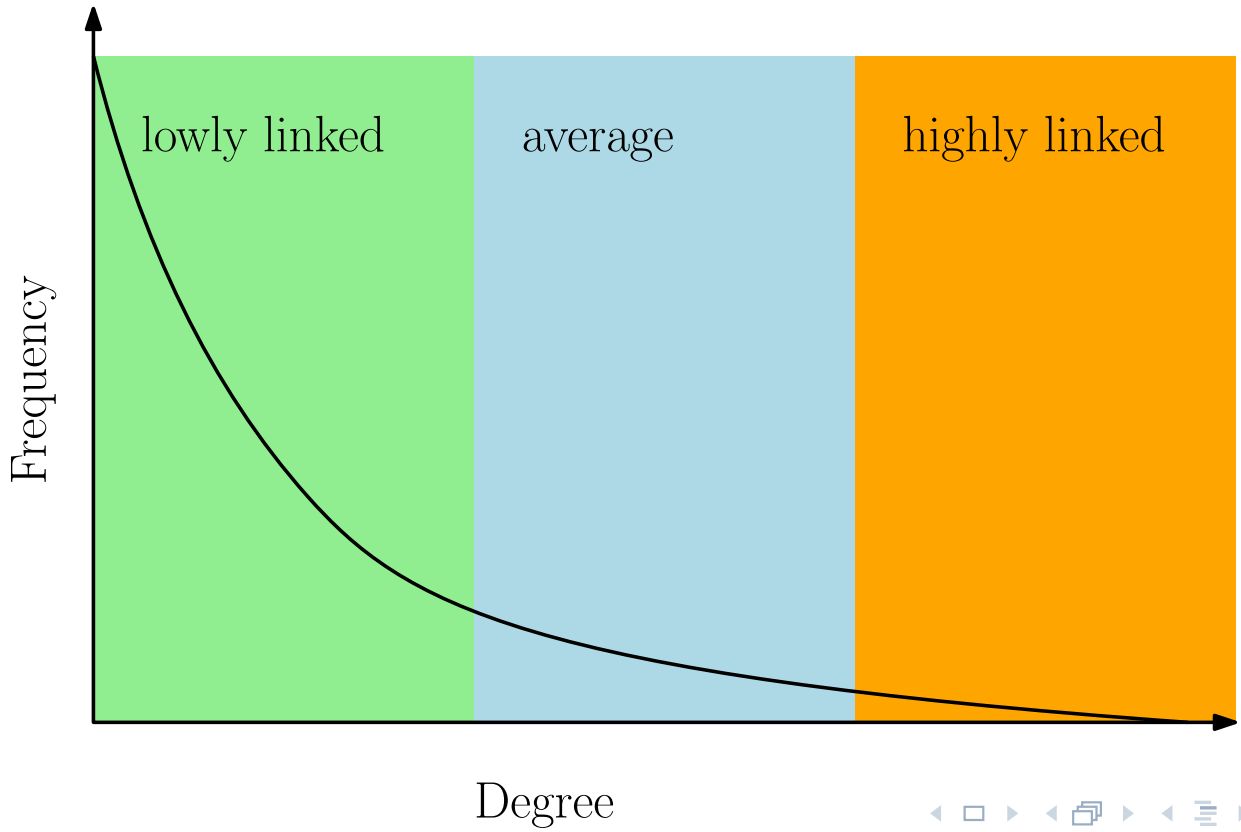
- In practice, we identify the graph G as *random graph* on V with edge probability p whose edge set is

$$E(G) = \{e | \omega(e) = 1_e\}$$

Definition 6.18 (Scale-Free Network)

A network is scale-free if the fraction of nodes with degree k follows a power law $k^{-\alpha}$, where $\alpha > 1$.

- Let $\mathcal{G}$ be the set of connected simple graphs with N nodes and $n(G) = (n_1, ..., n_{N-1})$, $n_k \in \mathbb{Z}$ be the degree histogram of $G \in \mathcal{G}$. This means, n_k is the number of nodes with degree k.



Definition 6.19

If $p = (p_1, ..., p_{N-1})$ with $p_k \in [0, 1]$ and $\sum_{k=1}^{N-1} p_k = 1$, then p has a power-law tail if $p_k = c(k - d)^{-\gamma}$ for $k > m$, where $d, m \in \mathbb{Z}, c, \gamma \in \mathbb{R}, d, m \geq 0$ and $\gamma > 1$.

Definition 6.20

$G \in \mathcal{G}$ is a scale-free graph if $n(G) \approx Np$ where p has a power-law tail.

What approximately equal means is not always clear.

Algorithm 11 Bollobás et all. (2003)

Require: $\alpha, \beta, \gamma, \delta_{in}, \delta_{out} \in \mathbb{R}^+$
Require: $\alpha + \beta + \gamma = 1$
Require: any fixed initial directed graph G_0 with t_0 edges

```
t := 0
while N(G_t) < n do
    r := random()
    if r < α then
        addNode(v)
        w = chooseNode(d_in, δ_in)
    else if r < α + β then
        v = chooseNode(d_out, δ_in)
        w = chooseNode(d_in, δ_in)
    else
        v = chooseNode(d_out, δ_out)
        addNode(w)
    addEdge(u, w)
    t := t + 1
```

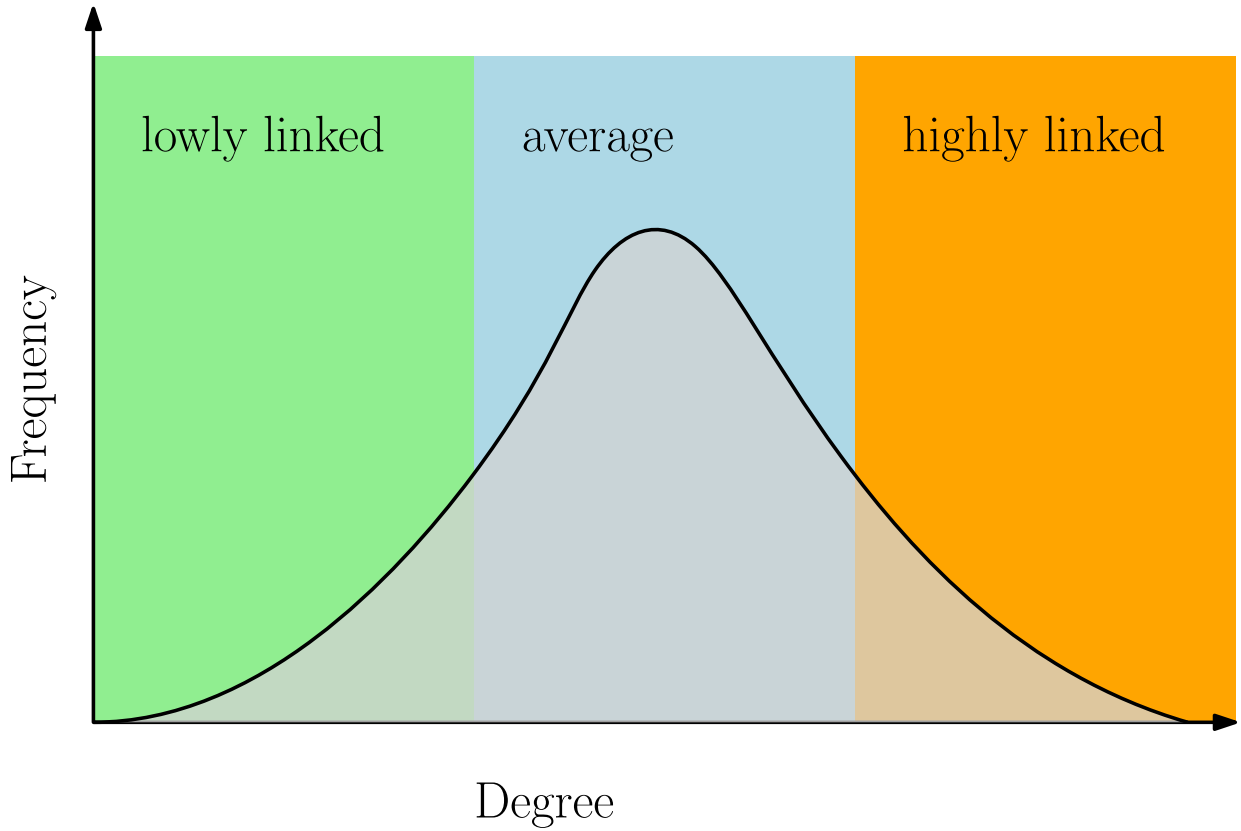
Algorithm 12 Bollobás et all. (2003)

1: Function{choose_node}{candidates, node_list, delta)}
2: Exercise

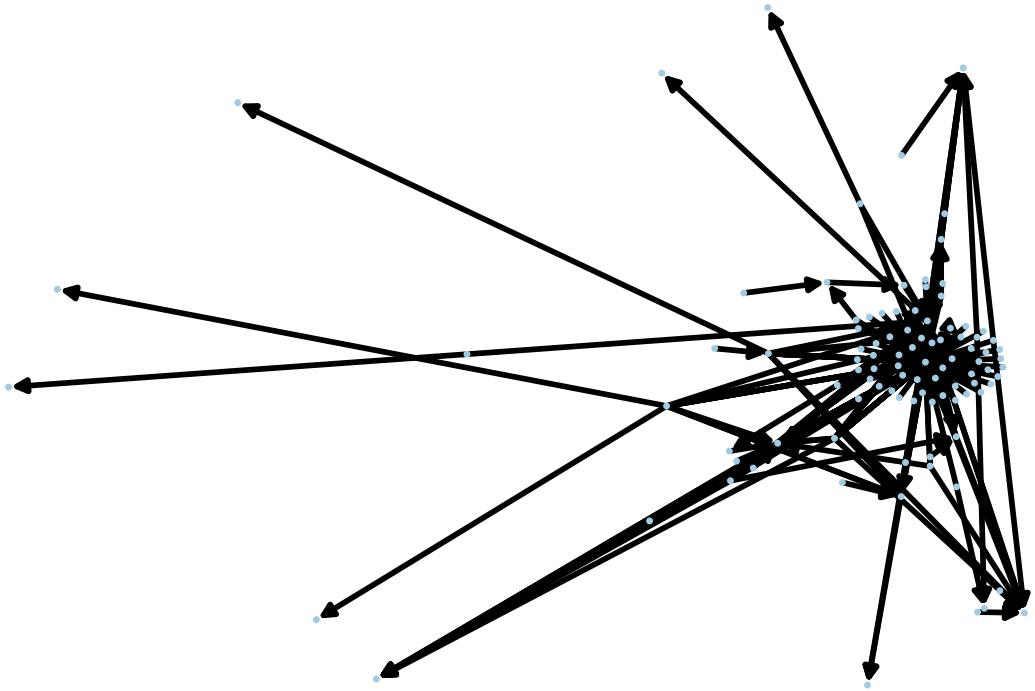
Definition 6.21 (Small World Network, Watts 1999)

Let $G = (V, E)$ be a connected graph with n nodes and average node degree k. Then G is a small-world network if $k \ll n$ and $k \gg 1$.

- Small world networks lead to high clustering, but short path lengths.
- Gaussian bell curve:



Random graph with $n = 100$ nodes.



Random graph with $n = 100$ nodes.