Baltic J. Modern Computing, Vol. 7 (2019), No. 2, pp. 271–292
https://doi.org/10.22364/bjmc.2019.7.2.06

About Correctness of Graph-Based Social Network Analysis

Mārtiņš OPMANIS

Institute of Mathematics and Computer Science, University of Latvia
Rainis blvd. 29, Riga, LV1459, Latvia
martins.opmanis@lumii.lv

Abstract. As a branch of network science, social network analysis widely uses graph techniques. Only in rare cases are results obtained from the graph models validated against "ground truth" and are directly applicable to objects in the investigated domain. Like extraneous solutions in mathematics, ungrounded mechanistic analogies, incorrect interpretation of indirect ties for intransitive relations and use of the "path" concept for social networks may lead to noninvertible results with no evidence outside the used graph model. The author investigates unimodal networks with dyadic ties, provides several examples of correct and incorrect applications and recovers the roots of incorrectness.

Keywords: Graphs, social network analysis, correctness, social experiment.

1 Introduction

Together with physical networks like transportation and computer-related networks, social networks comprising *actors* (humans or human-based structures like companies, parties, and social groups) and *relationships* (ties, interactions) between them are also investigated via attributed graphs. Excellent general overview of the history of graph usage in social network analysis is given in (Scott, 2002), while (Wasserman and Faust, 1994) contains an in-depth analysis and description of graphs in network analysis.

In this paper, unimodal networks with dyadic ties of single type among them will be investigated. An example of such social network is depicted in Figure 1.

We will maintain a clear line between *network* as a real-world phenomenon and its model — *graph*. The term "graph" here is used in the narrow sense of the word exclusively in connection with graph theory and has nothing with things

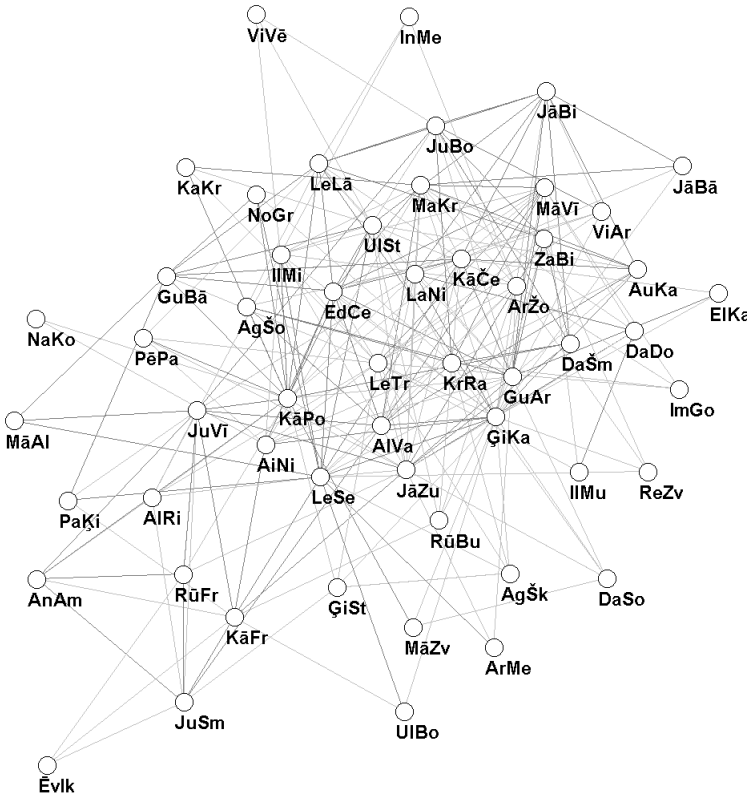

Fig. 1. An attributed graph of a social network.

like infographics, charts, and functions. It is assumed that relationships in the network should be verifiable without using any model including graphs.

Such a division is not obvious, since many authors use network and graph terms interchangeably: "My apologies here for the mixed terminology: *edge* and *node* are from graph theory; *tie* and *actor* are social network terms. You will need to be familiar with both usages, and I will use them interchangeably" (Robins, 2015). In others network terms are simply given as "synonyms" of graph terms (Bothorel et al., 2015): "Actor: also called a node or a vertex" (Denny, 2014), "... the propagation of a sexually-transmitted disease that spreads along the edges of a graph" (Watts and Strogatz, 1998), "Most often, nodes are individuals, such as individual persons or chimpanzees" (Borgatti et al., 2013).

Despite the fact that such interviewing justifies the naturalness of graph concepts for network analysis, it puts the reader under the delusion that **all** graph and network concepts can be used interchangeably and obtained results applied to the initial network in a simple and straightforward way.

We will divide the process of network analysis using graphs into three separate steps as schematically depicted in Fig. 2:

- $\mathcal{N}$ — obtaining an attributed graph from the real-life **n**etwork
- $\mathcal{A}$ — performing **a**nalysis on the graph
- $\mathcal{C}$ — applying analysis results and **c**onclusions from graph back to the network

Social network analysts follow this schema, usually without clearly subdividing the whole process into separate steps. If network and graph terms are used interchangeably, this gives the illusion that step $\mathcal{N}$ is not necessary and step $\mathcal{A}$ is (or can be) performed on the entities of the initial network. However, analysis is **always** based on the graphs and so any existing approach should be easily transferable to the described three-step schema even if it seems unnecessary puristic. In general, the same schema "create model — analyze model — apply results to the network" can be used also if a different network model is chosen instead of graphs.

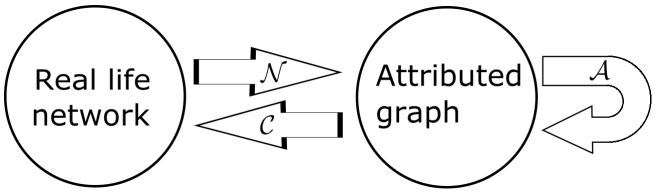

Fig. 2. The process of network analysis using graphs: $\mathcal{N}$ — obtaining an attributed graph, $\mathcal{A}$ — performing analysis, $\mathcal{C}$ — applying analysis results

In this paper, we will assume that the first two steps $\mathcal{N}$ and $\mathcal{A}$ are processed correctly and are completed, i. e. all known information (and nothing else) from the network is correctly transferred to the graph and all operations within the graph are performed in strong correspondence with graph theory.

This assumption is essential, since the literature contains mentions of several sources of incorrectness of these steps. For example, speaking about social networking services: "Unfortunately, many members of these sites try to connect with as many people as possible — whether they know them or not. This creates many false links/connections in the LinkedIn and Facebook databases. Two people might show to be connected, but they really are not — one person was too embarrassed to turn down a "friend request" from a total stranger" (Krebs, 2008). There may also be attempts to "enrich" data by adding ties that are not observed, since "it is wiser to look for more relaxed structures" (Bothorel et al., 2015 an introduction of quasi-cliques).

The main focus of the paper will be on the step $\mathcal{C}$ of applying graph results back to the network, since "The main goal of social network analysis is detecting and interpreting patterns of social ties among actors" (de Nooy et al., 2012).

Attention to the correctness of this step in the literature of social network analysis is surprisingly low. Just a few authors (Krebs, 2002, Kleinfeld, 2001) emphasize the necessity to validate results obtained from graphs with respect to the original network. The value of investigating network structure in isolation is also disputed: "More generally, the experimental approach adopted here suggests that empirically observed network structure can only be meaningfully interpreted in light of the actions, strategies, and even perceptions of the individuals embedded in the network: Network structure alone is not everything" (Dodds et al., 2003).

We can found similar critical thoughts aimed at inappropriate usage of numbers in general: "Numbers have become so familiar that we no more worry about when and why we use them than we do about natural language. We have lost the warning bells in our head that remind us that we may be using numbers inappropriately. They have entered (and sometimes dominate) our language of thought" (Edmonds, 2004).

In this paper, we will demonstrate that concepts of "path" as a chain of consecutive ties and "connectivity", which are natural for graphs and have good analogs in substantial networks, are not **always** applicable to social networks, and it is easy to get wrong conclusions based on such models.

The paper is organized as follows. Section 2 gives a short insight in graph concepts, Section 3 describes the general process of building attributed graphs from real-life networks. In the following Sections 4,5,6 and 7 problems with indirect ties and the incorrect use of several concepts in social networks due to intransitivity of ties are discussed. Transmission of messages is analyzed in Section 8. Several examples are analyzed thoroughly in the Section 9. Conclusions are given in Section 10.

2 Beyond the basics of the graph theory

The author assumes that the reader is familiar with graph concepts (Diestel, 2017, Wasserman and Faust, 1994, de Nooy et al., 2012) but would like to recapitulate some important graph features from the viewpoint of graph theory.

Definition 1. A *graph* is defined by two sets: set V of objects from some domain and set E of object pairs (v_1, v_2), where $v_1, v_2 \in V$.

Elements of V are called *vertices* or *nodes*, while elements of E are called *arcs* (if the order of objects in pairs is important) or *edges* (if the order is not important).

A particular graph by definition is **static** structure: V and E are fixed, and "analysis of the graph" means analyzing these two sets. In particular, it follows that graph models of dynamic networks can only be snapshots at particular moments of time or describe an underlying static structure.

The graph itself doesn't contain "historical" information on how sets V and E were created and why these sets contain exactly these elements. The "meaning" of V and E is out of scope from the viewpoint of the defined graph. Therefore, if there is any intention to apply results obtained from the graph to the initial

network, this meaning should be somehow kept beside the bare graph that is sufficient for graph-based analysis. The simplest approach is to add attributes (or properties (Needham and Hodler, 2019)) to the vertices and/or edges, like labels are added to the vertices in the graph depicted in Fig. 3. During graph analysis, labels or other attributes do not play any role and are used just to keep a backward connection between graph and the initial network.

However, simple labeling may be useless (like in Figure 1) if the reader is not familiar with the described domain and the labels are too weak for a proper "decoding". Let's investigate one more example.

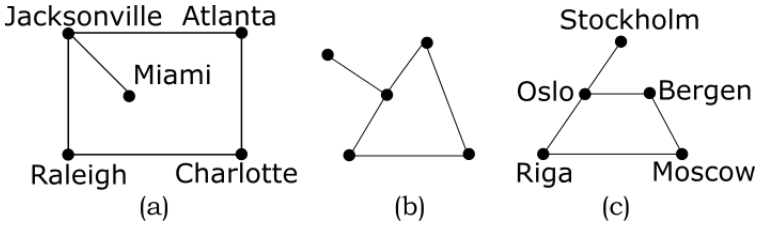

Fig. 3. Isomorphic graphs.

In the example depicted in Figure 3 three graphs are isomorphic, and since only structural relations matter, properties of all three graphs are the same. Graph (b) may be used for the analysis of graph properties, while the texts adjacent to vertices in (a) and (c) may be used just for back-referencing to the corresponding network. Judging from the names, some networks of cities from the USA (a) and Europe (c) are depicted in these graphs.

Assuming that (a) and (c) depict real networks, let's focus on them and try to answer the following questions:

- Since names of cities are given, is it possible to determine what networks are depicted in the corresponding graphs?
- Are relationships between cities in both networks the same?
- Are Jacksonville and Raleigh connected in the same way as their structural analogs Oslo and Riga?

Most probably it will be impossible to give certain answers to these questions without additional information. If we add information that in the (a) the relationship means "is connected by railroad", it becomes possible to give partial answer to the first question: "In (a) a small fragment of the USA's railroad network is depicted" as well as give negative answers to the last two questions since Riga and Oslo are not connected by railroad and therefore the relationship in (c) obviously differs from that in (a).

However, this knowledge gives nothing to recover the relationships in (c), while structural symmetry still encourages to draw parallels with (a). We will return to this example in Section 5.

So the overall conclusion is straightforward: the graph alone **cannot** be used to judge about the initial network if we do not know network details — what objects and relationships are depicted.

3 From network to graph

Let's investigate a simple example of how a graph can be obtained from a particular network. Let's try to describe a set of *movies*, assuming that each movie consists of several *episodes* and each *actor* of a particular movie performs in at least one episode. Our goal will be the investigation of collaborative work of movie actors and we will be interested in relationships of the form "Actors X and Y performed together in the same episode". What is the most appropriate way to build the corresponding graph?

The usual way is to define a single vertex for each actor and provide an edge for each appearance in an episode together. If information about all movies is collected together, discarding information on which movie each collaboration took place in, we can get a graph like in Fig. 4 (a), where appearing in the same episode in some movie for six actors A, B, C, D, E and F is shown. An edge between any pair of vertices denotes that corresponding actors performed together at least in one episode of at least one movie.

However, if we use multi-layer graphs (Kim and Lee, 2015) and focus on separate movies, we can create a separate graph (or graph "on a separate sheet") for each movie, like in Fig. 4 (b). Movie M_1 featured actors A, B, C, E and F, while M_2 featured B, C, D, E and F. As a result, there can be several vertices representing the same person in different movies.

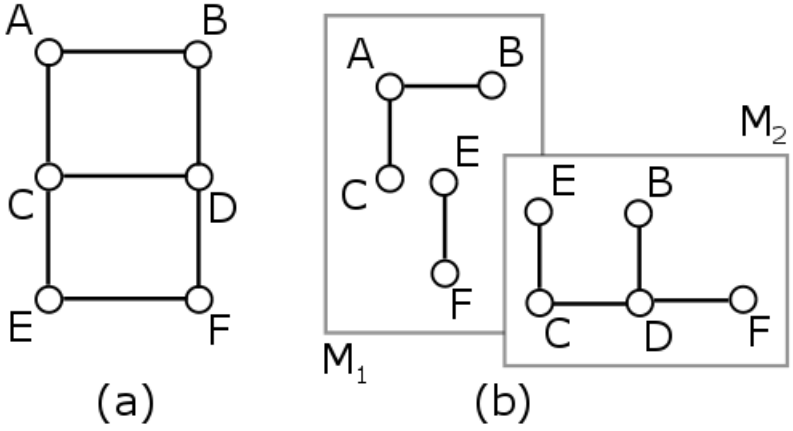

Fig. 4. Graphs obtained from the same network: (a) a simple graph neglecting particular movie information, (b) a multi-layer graph with a separate graph for each movie.

It should be pointed out that facts obtained from the network and depicted are the same for both graphs. From the viewpoint of graph theory, both obtained graphs are correct (all actors are depicted as vertices and all appearances in the same episode are depicted as edges). Due to its simplicity, the "all-in-one" way of modeling is preferred by network analysts and other possible approaches are not investigated. However, conclusions obtained from the graphs can substantially differ depending on the chosen approach. In our example, the question "Have the actors X and Y ever performed in the same episode?" can be answered from both versions while "Have the actors X and Y ever performed in the same movie?" cannot be answered from 4 (a) if there is no edge between the corresponding vertices. In particular, the answer is "yes" for D and E but "no" for their structural analogs D and A.

The choice of graph model highly depends on the research purpose. For example, if the intention is to investigate pairwise collaboration for a particular actor, an expressive characteristic of each vertex (*ego*) is obtained by investigating its induced 1-step subgraph (referred to as *egonet*) (Akoglu et al., 2010). In the given example, the egonet with ego B can better be explored directly in the "all (collaborations) in one" graph (Fig. 4 (a)). To obtain the number of different partners B had in any episode, we need only calculate the degree of vertex B (2). Collection of the same information from the multi-layer graph needs some preprocessing, e. g. creation of a virtual vertex B' where all appearances of B are collected together.

4 Direct and indirect ties

For direct ties, there is a straightforward bidirectional correspondence between graph objects and real-life artifacts.

If there are three pairs of mutual friends A and B, B and C, C and D, then this can be depicted as a simple graph (see Fig. 5):

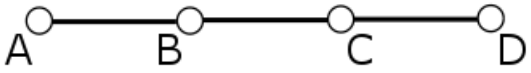

Fig. 5. A graph of three friendships.

If two persons are friends, then there will be an edge between the corresponding vertices, and there will be no edge if they are not. To discover whether two persons are friends, we look at the corresponding attributed graph of friendships, find the vertices marked by the persons' names and check whether there is an edge between them. As long as only direct ties are investigated, we can be sure that the graph corresponds to the real life. When building the graph, the relation "friendship" is assumed to be static: for a particular pair of persons it either takes place or not. It should be possible to verify this relationship without the

graph: if all involved persons X are asked "Is Y your friend?", their "yes/no" answers should match the previously gathered information.

But what can we say about indirect relationships between pairs of persons not tied directly, like A and D? Certainly, they are not friends (there is no edge between the corresponding vertices). Are they acquainted? Maybe yes (but then they are not friends) and maybe not — the relationship "is aquainted with" was not described in the initial set of facts for persons who are not friends and therefore is not presented in the graph regardless of how it is constructed. As a consequence, it is not possible to decide which answer is right without additional information about relationships besides friendship in the observed network.

The network of friends is a popular standard example, and several authors speak about "transitivity of friendship" in terms such as "it is a tendency for friends of friends to be friends" (Denny, 2014) or "the enemy of my enemy is my friend" (Borgatti et al., 2013 p.22). In real examples, "a friend of a friend is a friend" may be true "with high probability" (Hoff et al., 2002) but far from always.

The author himself is involved in friendly relationships with several people while direct relationships among these people are close to "being enemies" excluding any "friendliness".

In general, *any* assertion about relationships between persons not tied directly (such as A and D in Fig. 5) is just an *assumption* that cannot be justified from the given data.

5 Path concept

Let's continue with a few more concepts from the graph theory.

Definition 2. A *path* connecting two vertices u and v is an edge between them or a chain of consecutive edges via other vertices starting in u and ending in v.

The path is a natural concept for graphs. "Finding shortest paths is probably the most frequent task performed with graph algorithms and is a precursor for several different types of analysis" (Needham and Hodler, 2019). Due to the graph abstraction, it is always possible to perform an arbitrary number of simple steps from a vertex to a neighbor vertex via edge. We also can count the steps performed.

Definition 3. The *length of a path* is the number of edges in this path.

Also, we can introduce the term "connectivity".

Definition 4. Two vertices *are connected* if there exists a path between them.

Definition 5. The *distance* between two vertices is the length of the shortest path connecting these vertices or ∞ if the vertices are not connected.

Definition 6. A *connected component* is such a subset of vertices in an undirected graph that there is a path between any two vertices in this subset. There is no vertex outside this subset that has an edge to any vertex within the subset. An isolated vertex is also a connected component.

Definition 7. A *clique* is a subset of vertices in an undirected graph such that there is an edge between every two distinct vertices in this subset. There is no vertex outside this subset that has edges to all the vertices within the subset. An isolated vertex is also a clique.

Cliques together with *n-chains* (i. e. paths of length n) are introduced in the paper investigating group structures in social networks (Luce and Perry, 1949).

Connectivity in graphs as well as the use of terms "walk", "trail", "path" (Bondy and Murty, 1976 p.12) is so intrinsic that social network analysts neglect the necessity to define the corresponding constructs in the investigated domain and take for granted their meaningful existence. In (Peay, 1980), the necessity to choose the right approach to characterize connectedness for indirect ties is discussed, but still without questioning the correctness of the concept in general.

6 Relationships in a graph-based social network analysis model

Let us divide the class of all graphs into two disjoint classes: graphs where each connected component is a clique ($\mathcal{S}_C$) and all other graphs ($\mathcal{S}_N$). Representatives of these classes are depicted in Fig. 6.

A typical social network model is a graph where it is possible to find a connected component that is not a clique and therefore belongs to $\mathcal{S}_N$.

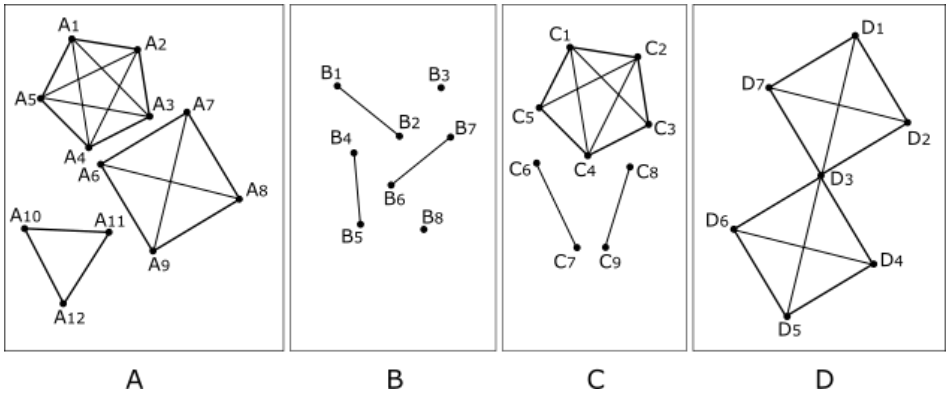

Fig. 6. Representatives of $\mathcal{S}_C$ (A, B) and $\mathcal{S}_N$ (C, D).

Non-completeness of at least one component is based on the observation that in real networks perfect structures are rare: "However, large cliques are difficult to find in real data because it is sufficient for one edge not to be present to break the clique, and in social graphs edges can be missing for many reasons, e. g., because of unreported data or just because even in a tight group there can be two individuals that do not get well together" (Bothorel et al., 2015).

Similarly, "Those nodes whose neighbors are very well connected (near-cliques) or not connected (stars) turn out to be "strange": in most social networks, friends of friends are often friends, but either extreme (clique/star) is suspicious" (Akoglu et al., 2010). And, "Obviously, social networks are neither complete not one-dimensional" (Fibich, 2017).

If there are separate connected components, they could be investigated separately (Banerjee et al., 2014). In case of a few outliers, attention is focused on the main group, excluding outliers from further analysis. Also, the opposite is possible, when researchers specifically look for anomalies in the graphs (Akoglu et al., 2015).

Definition 8. A binary relation R over a set of objects O is *transitive* if for any three objects $o_1, o_2, o_3 \in O$, $o_1 R o_2$ and $o_2 R o_3$ imply $o_1 R o_3$.

We demonstrated intransitivity of the friendship relation using an example in Section 4, but let us prove several propositions for two relations: E = "there exists an edge between two vertices" and P = "there exists a path between two vertices" for graphs from $\mathcal{S}_C$ and $\mathcal{S}_N$.

Proposition 1. Relation E over the set of all vertices of $g \in \mathcal{S}_C$ **is transitive.**

Proof. Since $g \in \mathcal{S}_C$, all connected components $c \subseteq g$ are cliques, so for any $v_i \in c$, $v_i E v_j$ and $v_j E v_k$ imply that also $v_j, v_k \in c$. Each vertex in a clique is connected with all other vertices in the clique. Therefore the transitivity requirement is fulfilled: $v_i E v_j$ and $v_j E v_k$ imply $v_i E v_k$. □

Proposition 2. Relation E over the set of all vertices of $g \in \mathcal{S}_N$ **is not transitive.**

Proof. Since $g \in \mathcal{S}_N$, there exists a connected component $c \subseteq g$ that is not a clique. There exist two vertices $v_x \in c$ and $v_y \in c$ that are not connected by an edge. Since c is connected, there exists a shortest path connecting v_x and v_y: $v_x E v_1, v_1 E v_2, \ldots, v_n E v_y$ with $n \geq 1$ intermediate vertices $v_1, v_2, \ldots, v_n \in c$. Let us look at any three consecutive vertices v_i, v_j and v_k on the path $v_x v_1 v_2 \ldots v_n v_y$. There is no edge between v_i and v_k — otherwise there exists a shorter path directly connecting v_i and v_k and not containing v_j. Since the given path is the shortest, this is impossible and we have found three vertices breaking the transitivity requirement: $v_i E v_j$ and $v_j E v_k$ do not imply $v_i E v_k$. □

Proposition 3. Relation P over the set of all vertices of $g \in \mathcal{S}_N$ **is transitive.**

Proof. By definition there are no vertices from distinct connected components having relation P. For any two vertices v_x and v_y from the same connected component, we have $v_x P v_y$. Therefore any three vertices v_x, v_y, v_z such that $v_x P v_y$ and $v_y P v_z$ belong to the same connected component and satisfy the transitivity requirement, as there is a path from v_x to v_z: $v_x P v_z$. □

Proposition 4. Relation P over the set of all vertices of $g \in \mathcal{S}_C$ **is transitive.**

Proof. The same as for **Proposition 3.**

These propositions show that in the case of $\mathcal{S}_N$ there is an important distinction between direct and indirect ties (or paths of length 1 and more than

1): direct ties **cannot be simply considered** a special case of longer paths, and paths do not automatically have the same features as direct ties. Features of indirect ties in the social network should be defined separately and they can not be simply deduced from the direct ones.

Now we return to the example depicted in Figure 3 (c) and reveal the secret that ties in this network are defined as "there exists a railroad connection between the cities **or** there is the same number of letters in the names of the cities written in English". The provided graph is formally correct but of low value for investigating indirect ties in the initial network of cities. It is not obvious that there is any valuable relationship between any pair of unconnected cities. Until such a relationship is defined (analogous to P in the theoretical model), it makes no sense to talk about anything based on the path concept.

Only when paths have a meaningful explanation, is it worth to calculate distances between vertices, seek shortest paths between pairs of vertices or calculate an overwhelming number of different graph *metrics* to analyze graph properties.

7 Roots of incorrect application of graphs

Questions about the correctness of representation almost never arise in physical networks: if roads are modeled, then it is possible to walk, run or drive on several roads in a row; and electric current can flow through several consecutive wires without a doubt. However, we can clearly see a difference between static structure (roads or wires) and dynamic processes that use this structure (someone walking or electric current flowing).

A usual way to explain social networks is to provide an analogy with the *static* structure of some physical network and further exploit the analogy of *dynamics* on an intuitive basis. Road or pipeline networks as well as electric circuits (Bozzo and Franceschet, 2013) are a few such analogs.

Borgatti et.al. in (Borgatti et al., 2013 p.3) discusses "interactions" forming "flows": "Flows may be intangibles, such as beliefs, attitudes, norms, and so on, that are passed from person to person. They can also consist of physical resources such as money or goods." Or, "Perhaps foremost among these is the idea that things often travel across the edges of a graph, moving from vertex to vertex in sequence — this could be a passenger taking a sequence of airline flights, a piece of information being passed from person to person in a social network, or a computer user or piece of software visiting a sequence of Web pages by following links" (Easley and Kleinberg, 2010). "Information flows" are also mentioned in (Krebs, 2008): "Employees who are included in key information flows and communities of knowledge are more dedicated and have a much higher rate of retention." In (Borgatti, 2005), "attitude influencing" and "emotional support" are mixed together with "e-mail broadcast" and "mitotic reproduction".

The semantics of the terms "walk", "trail", "path" assumes dynamics — that there is a possibility to "walk", "move" or "carry something" along a path. The term "flow" is also used with graphs (e. g. "maximum flow"), thereby assuming that there is something able to "flow", even if only as a quantitative abstraction.

Modeling networks by graphs implies a "possibility to travel" without limit via edges or chains of consecutive edges regardless number of already passed edges.

As in the foundational paper of graph theory, Euler's Bridges of Königsberg, we assume that a real person can cover distance necessary to cross all the bridges.

Topological distance metrics that are used for exploring social networks (like diameter, betweenness centrality, closeness centrality and eigenvector centrality) are based on the concept "path in a graph" (Hernández and Mieghem, 2011).

Physical networks may easy "blindfold" social network analysts if they hastily assume that social ties have the same characteristics as tangible ties in physical networks. The author insists that it makes a **substantial** difference whether in the original network there is a natural flow of things or a way to walk (money transfer, selling of goods, traveling of a particular person, surfing via links from one web page to the next) or the network is formed from static direct ties (friendship, having the same beliefs, conversations, asking for advice, e-mail communication, collaborative work) and there is no tangible and stable indirect flow between connected actors.

Although dynamic processes are justified for the physical networks (e. g. electric current can pass through several wires if they are connected), there are no general analogs for social networks!

Particularly interesting and confusing is the usage of the analogy of electric current when social ties "name of a person X is mentioned together with a name of a person Y on the same web page within a window of approximately ten words of one another" are investigated (Faloutsos et al., 2004). It is declared that there is some "current" from Alan Turing to Sharon Stone: "We note also that Alan Turing has direct connections to Alan Thicke, Alan Alda, and Bruce Lee (all of whom have direct connections to Sharon Stone), but these edges were discarded as **carrying too little current**." (emphasis mine). Of course, no evidence is given that there *exists* anything that can be regarded as *current* relevant to the real network and real people!

Since the 1950s, the term "social distance" (or "distance between individuals") has been used to describe a concept similar to "distance" in the corresponding graph (Bavelas, 1950, Scott, 2002 p.76, Kilduff and Krackhardt, 2008 p.69). This concept is explicitly based on the paths in a graph. It must be pointed out that back in 1967 S. Milgram already noticed a difference between "distance" in the real world and in a graph: "Almost anyone in the United States is but a few removes from the President, or from Nelson Rockefeller, but this is true only in terms of a particular mathematical viewpoint and does not, in any practical sense, integrate our lives with that of Nelson Rockefeller" (Milgram, 1967). Similar thoughts (with regard to graph diameter) can be found in (Denny, 2014): "A very large diameter means that even though there is **theoretically** a way for ties to connect any two actors through a series of intermediaries, **there is no guarantee** that they actually will be connected." (emphasis mine). Or in (Kleinfeld, 2001): "What does it actually mean in practical terms to be linked to others on a first-name basis? A welfare mother in New York might be connected to the president of the United States by a chain of fewer than six degrees:

Her caseworker might be on first-name terms with her department head who may know the mayor of Chicago who may know the president of the United States. But does this mean anything from the perspective of the welfare mother?" So there is no proof that there exist, and are usable, "paths" in the particular real networks!

8 Transmitting messages over networks

For some mental exercise, let us consider the relation "sends messages to" as described in (Luce and Perry, 1949) for two networks: a computer-based network with cables and communication devices like routers and switches and a human-based network that describes people with whom each person communicates, i. e. each person *is able to send* any message to any person from some list. Military structures transmitting orders are closer in this sense to the physical network since people *are obliged* to process information uniformly.

Despite the view "In the efficiency view of networks, the network simply operates as a passive conduit of information" (Carpenter et al., 2004), in a human-based network, there is no evidence that the initial message will be always passed in its original form through a long chain of actors. Of course, it can be done in an artificial environment, like in the movie "Six Degrees of Celebration" a concrete message from a particular child was carried to the president of Russia via social ties (Bekmambetov et al., 2010). Most probably we will get a "Chinese whispers" (Blackmore and Dawkins, 2000) game situation where the initial message will be lost in the chain of transmitting people. Even assuming that people are honest and willing to pass a correct piece of information, details are usually lost, added or transformed, making it almost impossible to recover in full detail the initial content of the message. Transmission of information is much more complicated, and several publications describe similarities in the spreading of epidemic diseases and of information (Goffman and Newill, 1967, Goffman, 1971). As pointed out in (Funk et al., 2009): "first-hand information about a disease case will lead to a much more determined reaction than information that has passed through many people before arriving at a given individual."

We can observe several factors that prevent messages from being carried over the network through a long chain of actors.

First, a message can survive only a limited number of transmissions: "... a new piece of information may only be "news" for a limited time. After while boredom sets in or some other news arrive and the topic of conversation changes" (Banerjee et al., 2014). In (Kadushin, 2012 p.206), a distance of three is mentioned as crucial: "Empirically, the influence of other persons or units on the focal person vastly declines somewhere between two and three steps out. It is not clear theoretically why this is true."

Second, there is a class of networks where it is impossible to reach a previously unknown addressee: "In a class of networks generated according to the model of Watts and Strogatz, we prove that there is no decentralized algorithm capable

of constructing paths of small expected length relative to the diameter of the underlying network)" (Kleinberg, 2000).

And, third, important factors determining whether a message will be carried may be hidden: "This may be because they are incorporating other information, such as who is trustworthy or who is most charismatic or talkative, which may not be picked up in the pure network data" (Banerjee et al., 2014). And, "This may seem counter-intuitive at first, but in fact, it formalizes a notion raised initially — in addition to having short paths, a network should contain latent structural cues that can be used to guide a message towards a target" (Kleinberg, 2000). In addition, information can be carried in disagreement with the physical laws in their mechanic analogs and alternative measures like flow betweenness (Newman, 2005) are invented.

The author has been able to find similar doubts in a few papers describing **real** experiments with the use of social ties (Travers and Milgram, 1969, Milgram, 1967). These tests have shown that there is an extremally high dropout rate: the number of completed chains is almost always under 30% (from 5% to 27.5%). Judith S. Kleinfeld found evidence that in S. Milgram's own other experiments the number of completed chains was even lower and this number highly depends on such real-life attributes as race and social class (Kleinfeld, 2002). In another later experiment (Dodds et al., 2003) number of completed chains was only 384 out of 24163 (1.59%). In the excellent overview of empirical small-world studies S.Schnettler shows that there are known just 11 serious real experiments from 1969 till 2003 all with very high drop-out rate (Schnettler, 2009).

An excellent conclusion is given in (Newman, 2005): "And even in a case such as the famous small-world experiment of Milgram (Milgram, 1967) and Travers and Milgram (Travers and Milgram, 1969), or its modern-day equivalent (Dodds et al., 2003), in which participants are explicitly instructed to get a message to a target by the most direct route possible, **there is no evidence that people are especially successful in this task** (emphasis mine)."

9 Examples

In this section, the author will provide several examples of graphs and possible conclusions obtained from them. It is easy to find examples where there is a natural and quite obvious meaning for indirect ties: the graph of citations (where vertices represent scientific publications, arcs show the relation "is cited in", and paths mean "is influenced by"), the World Wide Web (where vertices represent pages or separate resources, arcs show the relation "is linked to", and paths mean "is reachable from") are several examples of such networks with directed relationships. It should be pointed out that, while these networks represent social phenomena, they still are quite tangible.

9.1 Geospatial Network Model of the Roman World

An excellent representative of a correct model is ORBIS, The Stanford Geospatial Network Model of the Roman World (Scheidel et al., 2014), where the road

map of the Roman Empire can be investigated looking for shortest, fastest or cheapest routes. Various interesting results can be obtained through calculation and simulation. Since the modeled network is a physical network of roads, it is not surprising that it fits well in the world of graphs and there is a quite obvious one-to-one correspondence between network and graph constructs, and there is no doubt that calculations performed on the graph are compatible with the original network.

9.2 Consanguinity

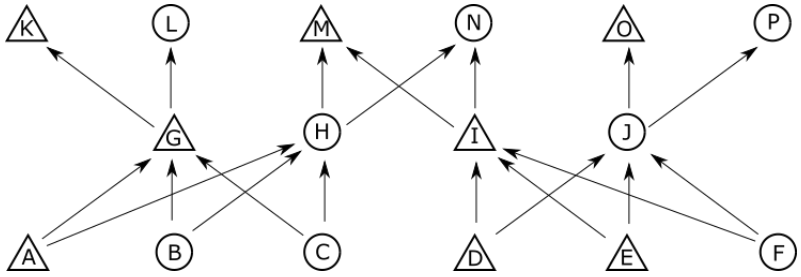

Fig. 7. A graph of a consanguinity network.

The next example is a graph of consanguinity where the depicted network consists of people tied with an "is a child of" relationship. Consanguinity is defined as "being related to someone by birth" or "having a common ancestor". An example of such a graph is given in Figure 7, where females are marked by rings, males by triangles and parents are placed above children. Usually, consanguinity relations are investigated from a particular person's perspective and it is possible to determine the *degree of kinship* as the length of a path that first goes upwards (from child to parent) and then goes downwards (from parent to child). Either of these parts may be absent, but they cannot be interchanged. With this restriction in place, the distance (or degree) between people in this graph is measured in a way which completely corresponds to Definition 5. For example, from A's perspective, degree 1 corresponds to the parents G and H, degree 2 to the grandparents K, L, M and N and siblings B and C, degree 3 to the uncle I, and degree 4 to the cousins D, E and F.

It should be pointed out that if we were to calculate the length of an arbitrary path without the above restriction, we could conclude that two people (e. g. K and P) are connected in a way that may be completely wrong in terms of the original network if there is no common ancestor.

9.3 Movie actor collaboration

A popular example used in social network analysis is a movie actor collaboration network built using data from the Internet Movie Database (IMDb)

(Needham., 1998, Borenstein, 2016). This undirected graph is built by modeling actors as vertices and connecting them with an edge if the corresponding actors have performed in the same movie. The famous parlor game "Six Degrees of Kevin Bacon" (Fass et al., 1996, Collins and Chow, 1998) is based on these data.

The famous actor Sir Thomas Sean Connery performed in 1957 in the movie "Hell Drivers" together with Wilfrid Lawson and in 1999 in the movie "Entrapment" together with Catherine Zeta-Jones (Connery, 2016). The corresponding attributed graph is depicted in Fig. 8 a). Since W. Lawson and C. Zeta-Jones have never performed in the same movie, the "distance" between W. Lawson and C. Zeta-Jones according to the graph is 2.

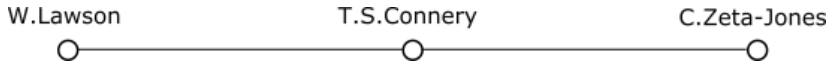

Fig. 8. Actor collaboration.

Traditionally, a finite distance (as opposed to infinite) is the sign that the particular objects are connected. However, W. Lawson passed away three years before C. Zeta-Jones was born (1966 and 1969 respectively), so there was no possibility in any tangible sense for C. Zeta-Jones to connect with and influence the non-existing W. Lawson.

Moreover, no existence or content of a possible "flow" between indirectly "connected" actors has been proven even for people who are alive.

9.4 Collaboration network and Erdős numbers

Another popular example is the network of joint publications (Borenstein, 2016). Each collaboration between a particular publication's coauthors, which constitute the basis of the built network, is correct: each vertex corresponds to a particular author, an edge between two vertices denotes a mutual publication and, most probably, also real collaborative work. A special case of the collaboration network is an attributed graph where "distance" from the famous mathematician Paul Erdős (1913–1996) is investigated (Easley and Kleinberg, 2010, Grossman, 2015). "Most mathematicians turn out to have rather small Erdős numbers, being typically two to five steps from Erdős. (...) The very existence of the Erdős number demonstrates that the scientific community forms a highly interconnected network in which all scientists are linked to each other through the papers they have written." (Barabasi and Frangos, 2014) The network is also mentioned in (Watts and Strogatz, 1998, Pelikán, 1996). The American Mathematical Society offers a free online tool to determine the Erdős number of any particular author (AMS, 2018).

Several sources give the impression that a smaller Erdős number is somehow related to a higher scientific value of a particular author. However, what **exactly**

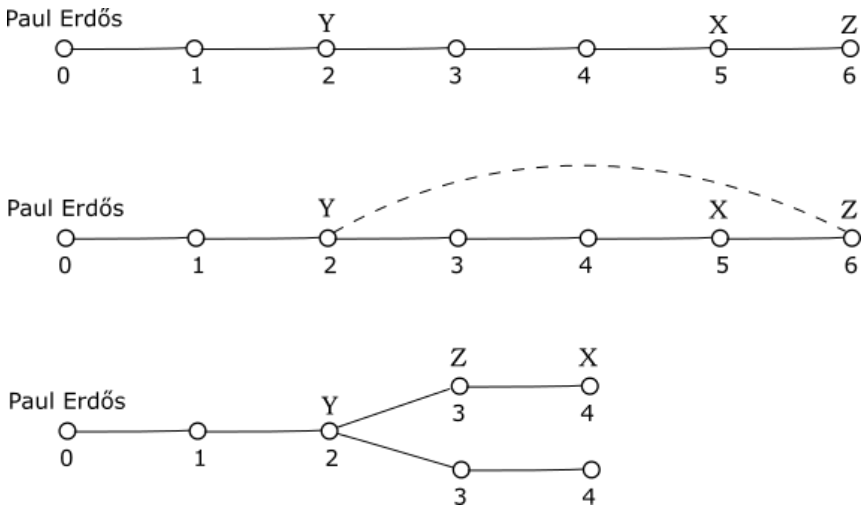

Fig. 9. Decreasing the Erdős number of X without direct involvement of X. a) initial state, b) new $Y - Z$ publication, c) updated state

is the meaning of "are linked through papers" for distances greater than 1, i. e. for persons who are not coauthors? Does having a lower Erdős number mean producing high-quality publications "by default"; is it enough to announce the Erdős number as proof of quality to make the author pass reviewing procedures and get published? Not at all. At least the author's personal experience shows that the same Erdős number may have authors with uncomparable scientific capacity.

An interesting justification that the Erdős number cannot be a stable measure of the "quality" of a particular scientist is the following (Ručevskis et al., 2016). It is possible to decrease the Erdős number of a particular author X without involvement of X: it is enough if some author Y on "X's social path to Erdős" publishes a paper with X's coauthor Z and as a consequence decreases also the Erdős number of X (see Fig. 9).

This is an essential difference from the consanguinity network described before, where relationships in the network are defined by birth and new relationships cannot be added without adding new actors.

If Erdős numbers cannot be considered an accurate measure of scientific quality, then is there any **meaning** to these numbers at all?

There may be the attempt to decide disciplinarity of publications from the collaboration network (Fortunato, 2010). If there are three authors being pairwise co-authors of some publication, then it can be decided that all authors are interested in the same subject. However, it is not always a case – as an counterexample, the author can name himself and two persons (Kārlis Čerāns and Juris Vīksna) having three pairwise connected publications (Viksna et al., 2007,

Opmanis and Čerāns, 2010, Čerāns and Vīksna, 1996) with content not related to the scientific interests of the third.

9.5 Paths to Putin

An excellent example of a network investigation that has almost all the dangerous features mentioned above is "Paths to Putin" by Valdis Krebs (Krebs, 2019). Using public data from journalists and court documents, a graph of more than 600 people and organizations is created and analyzed. Ties reflect business, political and/or personal contacts. It is declared that the gathered data demonstrates the existence of a covert relationship between the presidents of Russia and the USA, Vladimir Putin and Donald Trump.

The author would like to argue against some assertions made in this investigation.

- Citation: "When two individuals are trying to keep their relationships covert, they will never establish a direct tie between themselves." The paper uses the converse of this: since there were no verified meetings between Trump and Putin before the inauguration of Trump but there were meetings of people close to the presidents (so-called *associates*), this is treated as a sign of a covert relationship between the presidents. However, similar characteristics can be found in networks of representatives who are in a state of war: heads of countries have no direct contacts, while diplomats are usually constantly looking for chances to negotiate, and it is possible to find a shorter or longer chain of relationships going from one leader to another without any intention to have covert end-to-end communication. The individual relationships among associates may be caused by presidential orders, or even be a personal initiative using the president's name behind his back.
- Citation: "They will use trusted intermediaries to convey information/agreements or to pass money/resources between their two groups." Is there any proof that **anything** is passed via a chain of intermediaries from one president to the other? Most probably this is just an assumption based on another aforementioned fallacy: mixing up a verifiable resource (money) with an intangible one (information).
- Citation: "By running a simple network measure that looks at links within and between large groups, we found that both the Trump associates and Putin associates were linked mostly within their own group, but they had a significant number of ties to the other group! This implies that the ties between the two groups were probably not accidental, nor random." It should be pointed out that "network measure" here means "graph measure", and as long as there is no proven meaning to indirect ties, any attempt to claim that there is an obvious purpose or that the particular direct relationships are parts of one big plan and individuals work as a coordinated group in the original network is ungrounded.
- All conclusions are based on the "path" concept described earlier. Paths in the graph smoothly became "indirect paths from Trump to Putin or vice versa".

- Citation (emphasis mine): "Of course all of these paths are not used to communicate/conspire between the two sides, but it does give us an indicator what is **possible**." and "Of the 500+ **possible** paths between Trump and Putin, less than 20% were **probably** utilized, or attempted — and of those, less than 5% were **likely** to be relied upon." But is there any single *real* path for which this is demonstrably not just an assumption?
- Citation: "These 18 individuals are key in the network because they are on many paths of information flow — they know what flows." Of course *they* know, but is there any evidence that the origin of some bit of information that passes via an intermediate is one of the presidents and the other is its target?

10 Conclusions

Graphs are a powerful tool for the analysis of networks, and authors should themselves provide a critical evaluation of their choice of graphs as the model for a social network. However, usually such analysis is not provided or is based on wrong assumptions.

Usage of graphs cannot be admitted as correct if:

- Direct ties represent separate static facts, but reasoning assumes some unobserved dynamics.
- Analogy of intangible social ties with physical networks is declared without adequate explanation.
- A path concept is used for intransitive relationships, and graph metrics based on indirect ties are reflected back onto the original domain.
- There is no reasonable way to explain the internal meaning of numerical values and the phenomena observed in the graph in terms of the original network without circling back to graph concepts.

Any of the mentioned aspects should be a serious warning sign in the process of social network analysis and ask for careful revision of the used graph model.

Assumptions that social networks with intransitive relationships can be modeled in the same way as physical networks, along with graph metrics based on the concepts of path and connectivity via indirect ties, are the root causes of the observed problems.

If it is intended to go beyond ego and use graph metrics based on paths in graphs, it is crucial to verify the possibility to interpret indirect ties in the observed network and prove their transitivity.

Without any such proof, graph metrics based on path concept should not be used and conclusions based on indirect ties should not be made.

With the rise of machine learning, more and more effort should be put on validating the obtained results against the network. Routine translation of results back to the real life and basing other decisions on them without reasonable criticism is unacceptable.

Acknowledgements

This work was supported by ERDF project 1.1.1.1/16/A/135.

The author thanks Professor Kārlis Podnieks, Dr. Paulis Ķikusts and Oļegs Ošmjans for valuable comments.

References

Akoglu, L., McGlohon, M., Faloutsos, C. (2010). oddball: Spotting Anomalies in Weighted Graphs. In: *Advances in Knowledge Discovery and Data Mining*. 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part II. Springer Berlin Heidelberg, Berlin, Heidelberg, 410–421.

Akoglu, L., Tong, H., Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery* 29(3), 626–688.

American Mathematical Society (2018). MathSciNet - FreeTools. `https://mathscinet.ams.org/mathscinet/freeTools.html?version=2`

Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O. (2014). Gossip: Identifying Central Individuals in a Social Network. *NBER Working Paper No. 20422*, DOI: 10.3386/w20422. `https://www.nber.org/papers/w20422`

Barabasi, A., Frangos, J. (2014). *Linked: The New Science Of Networks Science Of Networks*. Basic Books.

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

Bekmambetov, T., Chevazhevskiy, Y., Jonynas, I., Kiselev, D., Voytinskiy, A. (2010). Movie "Six Degrees of Celebration" (original title – "Yolki"). `http://www.imdb.com/title/tt1782568/`

Blackmore, S., Dawkins, R. (2000). *The Meme Machine*. New ed edn. Oxford University Press.

Bondy, J.A., Murty, U.S.R. (1976). *Graph Theory With Applications*. Elsevier Science Publishing Co., Inc.

Borenstein, E. (2016). University of Washington course GS559: Introduction to Statistical and Computational Genomics (Winter 2016), Slides of lecture 15: Biological networks and Dijkstra's algorithm. `http://elbo.gs.washington.edu/courses/GS_559_16_wi/slides/15A-Networks_Dijkstra.pdf`

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

Borgatti, S.P., Everett, M.G., Johnson, J.C. (2013). *Analyzing Social Networks*. SAGE publications Ltd.

Bothorel, C., Cruz, J.D., Magani, M., Micenková, B. (2015). Clustering attributed graphs: Models, measures and methods. *Network Science* 3(3), 408–444.

Bozzo, E., Franceschet, M. (2013). Resistance distance, closeness, and betweenness. *Social Networks* 35(3), 460–469.

Carpenter, D.P., Esterling, K.M., Lazer, D.M.J. (2004). Friends, brokers, and transitivity: Who informs whom in washington politics? *Journal of Politics* 66(1), 224–246.

Collins, J.J., Chow, C.C. (1998). It's a small world. *Nature* 393, 409.

Connery, S. (2016). Filmography. `http://www.seanconnery.com/filmography/`

Čerāns, K., Vīksna, J. (1996): Deciding reachability for planar multi-polynomial systems. In Alur, R., Henzinger, Thomas A.and Sontag, E.D., eds.: *Hybrid Systems III: Verification and Control*. Springer Berlin Heidelberg, Berlin, Heidelberg, 389–400.

Denny, M. (2014). *Institute for Social Science Research, University of Massachusetts Amherst, Workshop "Social Network Analysis"* http://www.mjdenny.com/workshops/SN_Theory_I.pdf.

Diestel, R. (2017). *Graph Theory. Graduate Texts in Mathematics*. Springer-Verlag Berlin Heidelberg.

Dodds, P.S., Muhamad, R., Watts, D.J. (2003). An experimental study of search in global social networks. *Science* 301(5634), 827–829.

Easley, D., Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Edmonds, D.B. (2004). Against the inappropriate use of numerical representation in social simulation. http://bruce.edmonds.name/an/an.pdf

Faloutsos, C., McCurley, K.S., Tomkins, A. (2004). Fast discovery of connection subgraphs. In: Proceedings of the Tenth ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*. KDD '04, New York, NY, USA, ACM, 118–127.

Fass, C., Turtle, B., Ginelli, M. (1996). *Six Degrees of Kevin Bacon*. Plume.

Fibich, G. (2017). Diffusion of new products with recovering consumers. https://arxiv.org/abs/1701.01669v2

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* 486, 75–174.

Funk, S., Gilad, E., Watkins, C., Jansen, V.A.A. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences* 106(16), 6872–6877.

Goffman, W., Newill, V.A. (1967). Communication and Epidemic Processes. *Proceedings of the Royal Society of London Series A* 298, 316–334.

Goffman, W. (1971). A mathematical method for analyzing the growth of a scientific discipline. J. ACM 18(2), 173–185.

Grossman, J.W. (2015). The Erdős Number Project. https://oakland.edu/enp/

Hernández, J.M., Mieghem, P.V. (2011). Classification of graph metrics. *Technical report*, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 CD Delft.

Hoff, P.D., Raftery, A.E., Handcock, M.S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* 97(460), 1090–1098.

Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings*. 1 edn. Oxford University Press.

Kilduff, M., Krackhardt, D. (2008). *Interpersonal Networks in Organizations. Cognition, Personality, Dynamics, and Culture*. Cambridge University Press.

Kim, J., Lee, J.G. (2015). Community detection in multi-layer graphs: A survey. *SIGMOD Rec.* 44(3), 37–48.

Kleinberg, J. (2000). The small-world phenomenon: An algorithmic perspective. In: *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*. STOC '00, New York, NY, USA, ACM, 163–170.

Kleinfeld, J.S. (2001). Could It Be a Big World? http://www.judithkleinfeld.com/ar_bigworld.html

Kleinfeld, J.S. (2002). The Small World Problem. *Society* 39(2), 61–66.

Krebs, V.E. (2002). Social network analysis: An introduction by orgnet, llc. http://www.orgnet.com/sna.html

Krebs, V.E. (2008). Social capital: the key to success for the 21st century organization. IHRIM XII(5) 40

Krebs, V.E. (2019). Paths to putin. https://www.thenetworkthinkers.com (January 29, 2019)

Luce, R.D., Perry, A.D. (1949). A method of matrix analysis of group structure. *Psychometrika* 14(2), 95–116.

Milgram, S. (1967). The Small World Problem. *Psychology Today* 2, 60–67.

Needham, C. (1998). Internet movie database. textttthttp://www.imdb.com

Needham, M., Hodler, A.E. (2019) *Graph Algorithms. Practical Examples in Apache Spark and Neo4j*. OâĂŹReilly Media, Inc.

Newman, M.J. (2005). A measure of betweenness centrality based on random walks. *Social Networks* 27(1), 39–54.

de Nooy, W., Mrvar, A., Bategelj, V. (2012). *Exploratory Social Network Analysis with Pajek*. 2 edn. Cambridge University Press.

Opmanis, M., Čerāns, K. (2010). Multilevel data repository for ontological and metamodeling. In: *Databases and Information Systems VI-Selected Papers from the Ninth International Baltic Conference*, DB&IS.

Peay, E.R. (1980). Connectedness in a General Model for Valued Networks. *Social Networks* (2), 385–410.

Pelikán, J. (1996). Paul Erdős (1913-1996). *Mathematics Competitions* 9(2), 15–20.

Robins, G.L. (2015). *Doing Social Network Research: Network-based Research Design for Social Scientists*. 1 edn. SAGE publications Ltd.

Ručevskis, P., Podnieks, K., Kozlovičs, S., Grasmanis, M., Celms, E. (2016). Personal conversation.

Scheidel, W., Meeks, E., Grossner, K., Alvarez, N. (2014). Orbis – the stanford geospatial network model of the roman world. `http://orbis.stanford.edu/`

Schnettler, S. (2009). A small world on feet of clay? a comparison of empirical small-world studies against best-practice criteria. *Social Networks* 31(3), 179–189.

Scott, J. (Ed.) (2002). *Social Networks: Critical Concepts in Sociology*. Volume 1. Routledge.

Travers, J., Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry* 32(4), 425–443.

Viksna, J., Celms, E., Opmanis, M., Podnieks, K., Rucevskis, P., Zarins, A., Barrett, A., Neogi, S.G., Krestyaninova, M., McCarthy, M.I., Brazma, A., Sarkans, U. (2007). Passim – an open source software system for managing information in biomedical studies. *BMC Bioinformatics* 8(1), 1–7.

Wasserman, S., Faust, K. (1994). *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.

Watts, D.J., Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.

Received April 12, 2019 , revised May 5, 2019, accepted May 15, 22019