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. Social network analysis widely uses graph techniques. Together with correct applications, in some cases, results are obtained from the graphs using paths longer than one, and due to intransitivity of relationships, several metrics and results are not applicable backward to objects in the investigated domain in a meaningful way. The author provides several examples and tries to recover roots of an incorrect application of graphs.

Keywords: Graphs, social network analysis, correctness.

1 Introduction

In the process of building network models representation has vital importance: "Whether studying protein interactions, sexual networks, or computer systems, the appropriate choice of representation is key to getting the correct result." [1].

Attributed graphs together with sociometric and algebraic notation are a traditional form how to model networks [2]. The description of the network – famous bridges of Königsberg Leonhard Euler presented in the paper considered being the first paper in graph theory [3].

In this paper, we will draw a clear distinction between *network* as real world artifact and its model – *attributed (or labeled) graph.* Term "graph" here is used in strong connection with graph theory and has nothing with things like infographics, charts, and functions.

Graphs are based on just two concepts – *vertices* (or *nodes*) which can be connected by undirected *edges* or directed *arcs.* A pair of vertices may be connected by more than one edge or arc. By adding as attributes textual strings or numbers we obtain an expressive model of an investigated network – attributed graph. Definitions of various graph concepts can be found in [2,4]. If not given explicitly, the author will use graph terms in correspondence with [5].

In the case of physical networks (transportation and computer-related networks, electronic circuits and other tangible networks), the choice to use graphs as a source of analysis is determined by natural one-to-one correspondence between real life artifacts and graph constructs.

It is not surprising that there came idea to model in the same way real life objects: “The social networks have usually been formalized as graphs, and methods of graph theory have been used to motivate and organize the analysis.” [6]. Social networks comprise *actors* (humans or human-based structures like companies, parties, and social groups) and *relationships* (ties, interactions) between them. Excellent general overview of the history of graph usage in social network analysis is given in [7], while [2] contains in-depth analysis and description of graphs in network analysis.

However, concepts of “path” as a chain of consecutive edges or 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.

This paper describes author’s investigations and related general problems in social network analysis (SNA) with a focus on unimodal networks with people as actors and one type of dyadic ties among them.

The paper is organized as follows. Section 2 describes the general process of building social networks using attributed graphs. In the following three Sections 3,4 and 5 problems with indirect ties and incorrect use of several concepts in social networks due to intransitivity of ties are discussed. Two examples are analyzed thoroughly in the Section 6. Conclusions are described in Section 7.

2 General process of network analysis using graphs

The process of network analysis using graphs can be divided into three main steps:

□ $\mathcal{N}$ – obtaining an attributed graph from the real life **n**etwork
□ $\mathcal{A}$ – performing **a**nalysis on the created graph
□ $\mathcal{C}$ – applying analysis results and **c**onclusions back from graph objects to real life entities

and is schematically depicted in Fig. 1.

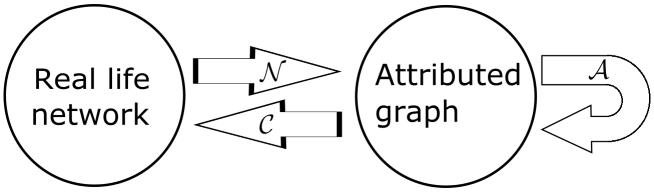

Fig. 1. Process of network analysis using graphs: $\mathcal{N}$ – obtaining attributed graph, $\mathcal{A}$ – performing analysis, $\mathcal{C}$ – applying analysis results

If we build a model of a road network, step $\mathcal{N}$ is building attributed graph (as a simpler version of the map without the geographical grounding of vertices) where each vertex denotes some city and each edge – road connecting a pair of cities. The attribute of a vertex usually is the name of the corresponding city, the attribute of an edge – a length of the corresponding road in kilometers. Step $\mathcal{A}$ is investigating traveling possibilities and providing estimations based on the graph like checking whether there is a route from one vertex to another or what is the shortest distance between them. Step $\mathcal{C}$ is making conclusions based on results of step $\mathcal{A}$ and possible acting – like choosing the best route for travel. An excellent representative of such model with different modes of road travel is ORBIS – The Stanford Geospatial Network Model of the Roman World [8].

It must be pointed out that word "attributed" in front of "graph" is essential. By losing attributes in step $\mathcal{N}$ – things like a name of a city or a length of a road in the previous example we obtain "bare" graph and usually lose the possibility to perform step $\mathcal{C}$ using just a graph structure as a set of anonymous vertices and edges.

Despite the truth that real life artifacts – actors and relationships between them are **not the same** as attributed graph concepts, in the literature networks and graphs are usually mixed up.

Network terms are given as "synonyms" of graph terms [9]. For example, "Actor: also called a node or a vertex" [10], "Most often, nodes are individuals, such as individual persons or chimpanzees." [11], "... the propagation of a sexually-transmitted disease that spreads along the edges of a graph." [12]. Term "small world" from real networks was transferred to graph analysis establishing a class of graphs – "small-world networks" [12] and there can be confusion whether network or graph is mentioned. Like, "The Small World problem" [13] (about networks) vs. "Could any graph be turned into a small-world?" [14] (about graphs).

Such examples of interviewing just demonstrate how naturally graph concepts fit in the minds of scientists, therefore, smashing differences of concepts, at the same time overlooking these differences in reality. From three mentioned steps, only $\mathcal{A}$ is out of the scope of this paper since it deals with clear graph constructs and all results obtained from this step are assumed to be correct. To ensure correctness of obtained results and conclusions regarding the real-life network, every transformation between real life and graph model (steps $\mathcal{N}$ and $\mathcal{C}$) **must be proven to be correct and meaningful**.

Correct representation of network data (performing step $\mathcal{N}$) is invaluable, especially if by data analysis decisions concerning particular people or society, in general, are made [15,16]. As Edward R Tufte says, "... there are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not. And, if the matter is an important one, then getting the displays of evidence right or wrong can possibly have momentous consequences." [15].

For social networks, it may be hard to verify collected ties and therefore ensure correctness of the whole network. Speaking about social networking services in [17]: "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."

As well there might be attempts to "enrich" data by adding ties which are not observed since "it is wiser to look for more relaxed structures" [9] (an introduction of quasi-cliques). Also, indirect falsification may take place after completing the step $\mathcal{N}$. In [14] is proven possibility to transform graph obtained from the social network to the "small-world" – a graph having different characteristics. It is not noticed, that transformed graph loses connection with the initial network since backward transformation will demand a change of real social ties, which seems at least strange. Since authors even not try to perform $\mathcal{C}$, such doubts do not arise in their paper.

If authors talk about differences between social and other networks, just quantitative differences are emphasized [18] without noticing essential differences. Step $\mathcal{N}$ may be more complex if hypergraph approach will be used in transfer process from a network to a graph [6].

The main focus of the paper will be on the last, and essential step $\mathcal{C}$ since "The main goal of social network analysis is detecting and interpreting patterns of social ties among actors." [4] Attention to step $\mathcal{C}$ in the SNA literature is surprisingly low. Just a few authors hold a view that step $\mathcal{C}$ is necessary even with correct step $\mathcal{N}$ since obtained $\mathcal{A}$ results themselves are not sufficient to judge about social network properly: "...maps and metrics are mirrors, not report cards! The consultant and the client together make sense of what the maps/metrics reflect about the organization." [19] And, "Such important work in mathematics begs for psychological research: What do people actually believe about the navigability of their social worlds and how do such beliefs influence their search attempts and their search success?" [20].

Pitfall also may be a nice visualization of attributed graph leaving $\mathcal{C}$ as a "homework" for readers usually misleading them and pushing to incorrect conclusions due to biased coloring schemes or by a geographical grouping of vertices giving "clues" how to "read" the graph.

3 Direct and indirect ties

For direct ties, there is a straightforward bi-directional correspondence between graph objects and real life artifacts and raising a question about correctness seems to be ridiculous. City X and a vertex corresponding to city X in a graph have a good mental connection. The same goes for a road between two cities and corresponding edge in the graph. If there is the edge between two vertices X and Y, then we can be sure that in real life road connects cities X and Y. So there is almost no difference if we speak about the connectivity of vertices in the graph or real cities and a connecting road.

The same situation is if direct ties from social networks are transformed to the graph. If two persons are friends, then there will be an edge between corresponding vertices, and there will be no edge if they are not. To discover whether two persons are friends we must take a look at the corresponding attributed graph of friendships, find two vertices marked by person's names and check whether there is an edge between them or not. So we can ascertain that graph corresponds to the real life as far only direct ties are investigated.

Besides single ties, it is possible to analyze also sets of such ties. Like in the [21] expressive characteristic of each vertex (an *ego*) is obtained by investigating its induced 1-step sub-graph (referred as *egonet*). In the context of the current paper, the egonet edges not incident with the ego should be treated of another kind.

For example, there may be the attempt to decide disciplinarity of publications from the collaboration network [22]. 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 example of the close scientific circle author can name himself and two persons having three pairwise connected publications [23,24,25] with content not related to the scientific interests of the third party.

The interest of social network researchers is not limited to direct ties in a network – also chains of consecutive ties are investigated. The reflection of any such chain in the graphs is *path*.

Definition. *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. We can perform series of simple steps from a vertex to a neighbor vertex, and there are no reasons why it would not be possible. We also can count steps performed.

Definition. *Length of a path* is number of its edges.

Also, we can introduce term "connectivity".

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

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

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

Definition. *Clique* is a subset of vertices in an undirected graph such that there is an edge between every two distinct vertices from this subset. There is no vertex outside this subset having edges with all vertices from the subset. An isolated vertex also is a clique.

Cliques together with *n-chains* (i.e. paths of length n) are introduced in the paper investigating group structures in social networks [26].

For a particular connected component, it is possible to also calculate *characteristic path length* as a length of the shortest path between two vertices, averaged over all pairs of vertices [12].

4 Relationships in a graph-based SNA model

A typical representative of the social network model is a undirected graph where it is possible to find a connected component which **is not** a clique. Let us denote the class of all such graphs as $\mathcal{S}$. Examples of graphs not belonging (A and B) and belonging (C and D) to $\mathcal{S}$ are given in Fig. 2. We did not insist that graphs A and B are not obtainable from the real networks – just that networks having all possible pairwise ties among actors inside all separated groups are not a subject of sophisticated analysis using graphs.

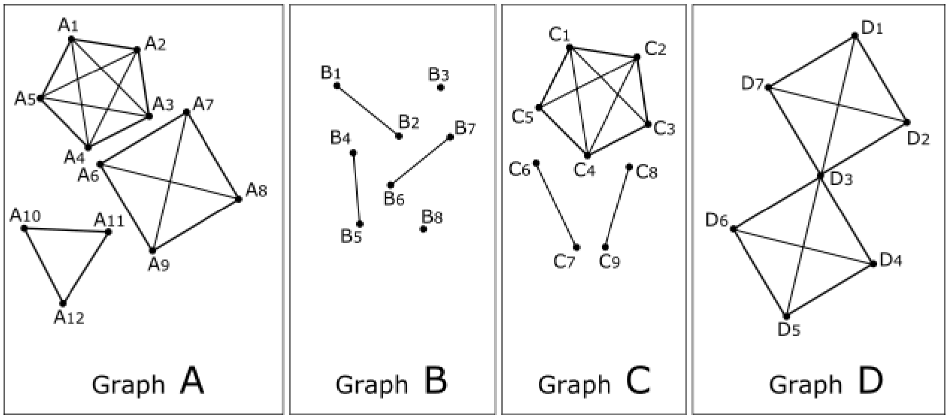

Fig. 2. Examples of graphs - A and B does not belong to $\mathcal{S}$, C and D belong to $\mathcal{S}$

Non-completeness of at least one component is based on the assumption 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." [9]. 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." [21]. And, "Obviously, social networks are neither complete not one-dimensional." [27].

If there are separate connected components, they are investigated separately [28]. In a case of few outliers, the focus is paid to the main group excluding outliers from the further analysis. Also, the opposite is possible – when researchers instead of investigating traditional patterns look for anomalies in the graphs [29].

Definition. 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 $o_2 R o_3$ implies $o_1 R o_3$.

Proposition 1. Relationship $E =$ "there exists an edge between two vertices" over the set of all $g \in \mathcal{S}$ vertices **is not transitive**.

Proof. Since $g \in \mathcal{S}$, there exists connected component $c \subseteq g$ being not clique. There exists two vertices $v_x \in c$ and $v_y \in c$ not connected by edge. Since c is connected, there exists shortest path connecting v_x and v_y: $v_x E v_1, v_1 E v_2, ..., v_n E v_y$

with $n(n \geq 1)$ intermediate vertices $v_1, v_2, ..., v_n \in c$. Let us look to any three consecutive vertices v_i, v_j and v_k on the path $v_x v_1 v_2 ... v_n v_y$. There is no edge between v_i and v_k – othervise there exists shorter path directly connecting v_i and v_k not containing v_j. Since given path is the shortest, this is impossible and we found three vertices breaking transitivity requirement: $v_i E v_j$ and $v_j E v_k$ does not imply $v_i E v_k$. □

Proposition 2. Relationship $P =$ "there exists a path between two vertices" over the set of all $g \in \mathcal{S}$ vertices **is transitive**.

Proof. By definition there are no vertices from a distinct connected components having relationship P. For any two vertices v_x and v_y from the same connected component takes place $v_x P v_y$. Therefore any three vertices v_x, v_y, v_z having $v_x P v_y$ and $v_y P v_z$ belongs to the same connected component and satisfy transitivity requirement since there exists path from v_x to v_z: $v_x P v_z$. □

The analogous propositions for directed graphs can be easily proven just by substituting edges by arcs.

Both Propositions show that there is the essential difference between direct and indirect ties (or paths having length 1 and greater than 1) – direct ties **can not be simply considered** as a special case of longer paths!

5 Roots of an incorrect application of graphs

5.1 Incorrect use of connectivity due to intransivity of ties

Connectivity in graphs as well as usage of terms "walk", "trail", "path" [30, p.12] is so intrinsic that social network analysts neglect the necessity to define corresponding constructs in the investigated domain and takes for granted meaningful existence of them also there. In [31] necessity to choose the right approach to characterize connectedness for indirect ties is discussed still not raising the question about the correctness of concept in general.

Semantics of terms "walk", "trail", "path" assumes that there is possibility to "walk", "move" or "carry something" via path. Also in graphs is used term "flow" (e.g. "maximum flow") assuming that there is something able to "flow" even as a quantitative abstraction. Graph abstraction itself implies possibility to "travel" via edges or chain of consecutive edges without limitations. Only in these circumstances, it is possible to calculate distances between vertices, seek for shortest paths between pairs of vertices and do similar things.

Questions about the correctness of representation almost never arose in physical networks - if roads are modeled, then it is possible to walk, run, ride using several roads in a row, electric current can pass several consecutive wires without a doubt. Physical networks "blindfold" SNA analysts and they overlooked this disagreement. In [11, p.3] is written about "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." [32]. "Information flows" are also mentioned in [17]: "Employees who are included in key information flows and communities of knowledge are more dedicated and have a much higher rate of retention."

There is essential difference whether in the original network there is natural flow of things or a way to walk (money transfer, selling of goods, travelling of a particular person, going via physical links from one web page to the next) or the network is formed from a 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. Usually, in publications about social network analysis authors hastily assume that social ties have the same characteristics as tangible ties. For example, in [33] "attitude influencing" and "emotional support" are mixed together with "e-mail broadcast" and "mitotic reproduction".

Particularly interesting is the attempt to use 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 [34]. 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, there is no given any evidence that there *exists* anything that can be counted as *current* relevant to the real network and real people!

Since the nineteen-fifties term "social distance" (or "distance between individuals") was used to describe concept similar to "distance" in the corresponding graph [35], [7, p.76], [36, p.69]. This concept explicitly is based on the paths in a graph. It must be pointed out, that back in 1967 S.Milgram already noticed 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." [37] The similar thoughts (when speaking about graph diameter) you can find in [10]: "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 [20]: "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 we are allowed to use “paths” in the particular real networks!

Therefore, despite connectivity in the corresponding graph, relations **may be not extendable to indirect ties** if direct ties in social networks reflect independent observations!

For example, the network of Padgett’s Florentine families includes the set of sixteen Italian families in the early XV century [2, p.103]. There is exploited symmetrical, but intransitive relation “a member of family X is married to a member of family Y” having no meaning for indirect ties.

A popular standard example is a network of friends, and several authors also speak about “transitivity of friendship” in terms “it is a tendency for friends of friends to be friends” [10] or “the enemy of my enemy is my friend” [11, p.22]. In real examples “friend of a friend is friend” may be “with high probability” [38] but far from taking place always.

Almost anyone reader from his experience from facts that X and Y are friends and Y and Z are friends can find examples with contrary results playing as Y. Three simple outcomes for a relationship between X and Z can be:

- □ X and Z are friends in real life, but this tie is not reflected in the graph,
- □ X and Z are representatives of non-overlapping domains of Y interests and are not familiar and therefore are not friends,
- □ a mutual relationship between X and Z is close to “being enemies” despite knowing each other perfectly.

These different outcomes show that there is no such thing as “friendship flow” going beyond direct ties and defines a relationship between X and Z in a deterministic way. In the literature, we can find examples where authors discuss this topic of intransitivity but at the same time use highly simplified approach assuming transitivity of friendship ties and investigates exclusively balanced networks, and not considering other cases [36, p.68].

It must be pointed out, that there are networks representing society which can be considered correct from the viewpoint of transitivity. Kinship graph (vertices represent persons, arcs – relation “is child of”, paths – “is descendant of”) used in investigation of spreading genetically grounded diseases, graph of citations (vertices represent scientific publications, arcs – relation “is cited in”, paths – relation “is influenced by”), World Wide Web (vertices represent pages or separate resources, arcs – relation “is linked to”, paths – “is reachable from”) are a few examples of such networks. In all mentioned examples ties or relationships are directed and should be modeled by arcs instead of edges.

As well physical nature of a network does not imply correctness of “path” concept. For the network of roads, if there are roads connecting cities A and B and also connecting cities B and C, then we can assume that we can also get from A to C passing B. It is usually true either for humans or vehicles. However, if instead of physical roads we investigate routes of public transportation how to get from A to C, then it is possible that at B we need to switch from train to bus if there is no train connection between B and C. In this case “it is possible to get from A to C” is still true for a particular human, but not for a particular train carrying passengers from A to B. So it is necessary **always** understand modeled network.

5.2 Transmitting messages over networks

As a good comparison may be used relation “sends messages to” already described in [26] for two networks: computer-based with cables and communication devices like routers and switches and human-based network which describes people with whom particular person communicates, i.e. person *is able to send* any message to any person from some list. Military structures and transmitting orders in this sense are closer to the computer-based network since people *are obliged* to process information uniformly. But even in computer-based networks not always message is carried to the right addressee via intermediaries due to packet loss and other technical problems.

Despite view “In the efficiency view of networks, the network simply operates as a passive conduit of information” [39], in a human-based network, there is no evidence that 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” the concrete message from a particular child was carried to the president of Russia via social ties [40]. Most probably we will get “Chinese whispers” [41] 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 usually are lost, added or transformed making almost impossible to recover in details the initial content of the message. Transmission of information is much more complicated, and in several publications, there is described similarity of spreading epidemic diseases and information [42,43]. As pointed out in [44]:”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.”

Against possibility that message may be carried over the network through a long chain of actors, works three observations.

First, any message can survive just 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.” [28]).

Second, there is a class of networks where it is impossible to reach 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).” [45].

And, third, important factors determining whether a message will be carried or not 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.” [28]. 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.” [45].

Similar doubts author can find only in the papers describing a few known **real** experiments with the usage of social ties [37,46]. These tests have shown that there is extremally high dropout rate – the number of completed chains almost always is under 30% (from 5% till 27.5%). Judith S. Kleinfeld had found evidence that in other S.Milgrams experiments the number of completed chains was even lower and this number highly depends on such real-life attributes as race and social class [13]. On a few experiments with a dramatically low success rate whole theory is built without further attempts to ground obtained results in real life!

Also in few more publications concerning possible pitfalls in social network analysis [18,47] authors oversee malformation in the foundation.

5.3 Misleading metrics and clusters

There is invented an overwhelming number of different graph *metrics* to analyze graph properties. A lot of them are also used for exploring social networks. All topological metrics of distance class (like diameter, betweenness centrality, closeness centrality and eigenvector centrality) are based on concept “path in a graph” [48]. However, according to the written above, such metrics **are not applicable** to the networks with intransitive relationships!

Surprisingly, there are attempts to apply some of these metrics to physical networks using as a justification their successful use in SNA [49,50].

One of the popular ways to get information about networks in the concentrated form is to seek for communities what in the world of graphs means usage of clustering algorithms and obtaining *clusters*. Related (closely connected) objects are included in the same cluster and distinct (weakly connected) objects – in separate clusters and there are much more edges having endpoints from the same cluster if compared with edges with endpoints from the distinct clusters. Clusters may be non-overlapping (each vertex belongs to at most one cluster) or overlapping (a particular vertex may belong to several clusters at the same time). There are no uniform criteria for vertex division in clusters applicable for all cases. Besides simple and crystal-clear cases, like “all members of a particular connected component should belong to the same cluster”, there still is a space for several distinct approaches.

A group of clustering algorithms exploits walk-based approaches and use the already discussed concepts of “distance”, “walk” and “path”, while others “directly extend well-known and efficient graph-based methods” [9]. When there is no reasonable meaning of graph terms in the original network also task to find clusters **by this approach is meaningless**.

This paper was inspired by the social network depicted in Fig. 3 which represents historical data from the defense of students bachelor thesis at Faculty of Computing of the University of Latvia. Named vertices correspond to supervisors and reviewers (a particular person may have served in both roles). Arc corresponds to the particular thesis and goes from vertex A to vertex B if the first one corresponds to the supervisor and the second - to the reviewer of this thesis. There can be multiple arcs in both directions between the pair of vertices if there are several theses with the same pair of supervisor and reviewer.

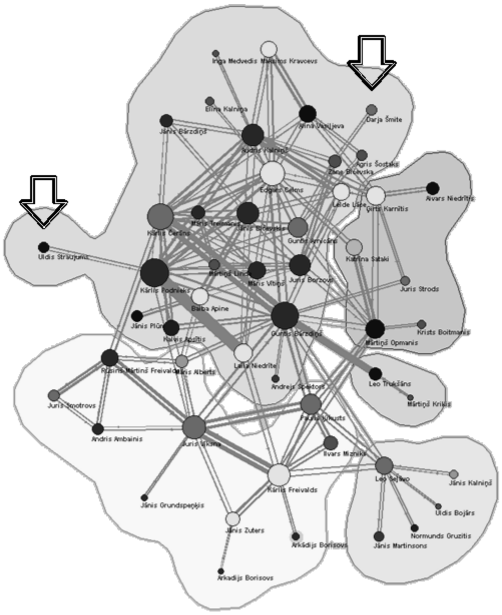

Fig. 3. Example of social network with clusters.

The picture with clusters given in Fig. 3 is an outcome of advanced graph visualization algorithm [51]. Depicted graph in the same cluster contains two vertices (depicted by arrows) corresponding to persons known by author and having not so much in common. However, authors of clusterization and visualization algorithms refused to discuss reasons for such placement in terms of network and were ready to talk only about graph constructs and algorithms themselves [52]. There is no doubt that graph without clusters has crystal-clear backward correspondence with real life data and if we limit ourselves to the direct ties, there are no problems with the correctness of the built graph. Regarding the general process of SNA, step $\mathcal{N}$ was completed correctly. There is nothing to argue also against $\mathcal{A}$, except observation that there is no transitivity of ties which in its turn lead to a useless effort in finding non-trivial clusters. Since obtained results have no sense in the network, there is no correct way to complete $\mathcal{C}$.

At the same time by just observing Fig. 3 we are pushed to the conclusion that these two persons (not vertices!) have "something" in common without additional evidence – just because corresponding vertices reside in the same colored area. However, such visualizations may be a good starting point for conjectures leading to a discovery of hidden characteristics of the network, which should be provable without graph models.

6 Examples of an incorrect application of graphs

6.1 Movie actor collaboration

The popular example used in SNA is movie actor collaboration network which is built using data from the Internet Movie Database (IMDb) [53,54]. This undirected graph is built modeling actors as vertices, and a particular edge connects two vertices if corresponding actors performed in the same movie. The famous parlor game "Six Degrees of Kevin Bacon" [55] is based on these data.

Let's investigate small example: Famous actor Sir Thomas Sean Connery in 1957 performed in the movie "Hell Drivers" together with Wilfrid Lawson and in 1999 in the movie "Entrapment" together with Catherine Zeta-Jones [56]. The corresponding attributed graph is depicted in Fig. 4 a). In the proposed model ties are undirected, and their meaning is just in explaining a fact that particular two actors performed in the same movie. If we look at this graph just as the collection of static facts (edge denotes just performing in the same movie and **nothing else**) then the graph is correct "snapshot" of the history, while the amount of information deducible from the graph is limited to investigating **just direct links**. It is possible to get a total number of movies where the corresponding actor performed (calculate degree centrality) or find the number of appearances of the pair of actors acting together in a movie (get a weight of the particular multi-edge). Such observations and calculations based only on direct ties are correct (but not so interesting since can be simply obtained without graphs).

However, SNA investigators usually do not stop there and start to calculate distances between movie actors who never performed in the same movie. Since W.Lawson and C.Zeta-Jones never performed in the same movie, distance in the one-mode network between corresponding vertices of W.Lawson and C.Zeta-Jones by definition is 2. Behind the scenes, we can feel the attempt to put impression that distance is not between vertices, but real persons!

Assuming, that this is so, there must be something allowing to connect W.Lawson and C.Zeta-Jones. To speak about distances greater than 1, we should imagine, that network instead of the simple fact of appearance shows some unapproved "transfer" (or "flow") of intangible things like memories, jokes, and attitudes. However, such transfer by its nature is directed (goes from the one collaborating actor to the other) and therefore each collaboration should be modeled by a pair of arcs. So we will obtain the modified graph Fig. 4 b). For direct ties this is acceptable – each actor may get "something passed via this link" from all other actors taking part in the particular movie (assuming that this is real collaborative *work* and not just appearing together in the movie credits).

However, assuming that "flow of something" can go beyond direct interaction, it may go from C.Zeta-Jones to W.Lawson as well as in the opposite direction (both paths are created by two consecutive direct "flows"). In real life, W.Lawson passed away three years before C.Zeta-Jones was born (1966 and 1969 respectively), so there was no possibility in any sense for W.Lawson to get "something" from non-existing C.Zeta-Jones. So just one of flows depicted

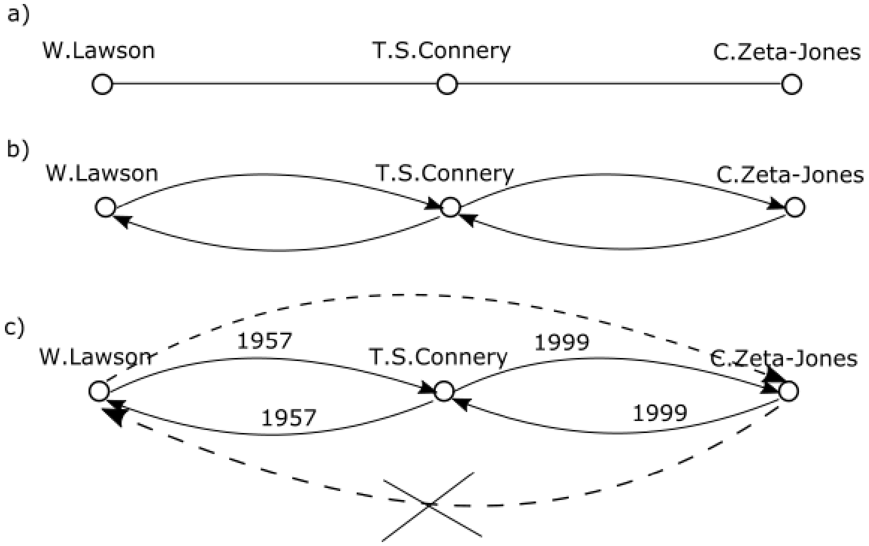

Fig. 4. Graphs of actor collaboration.

in Fig. 4 c) can take place, but this can not be deduced looking just on the graph where the corresponding vertices are symmetric to each other. One of the solutions is adding the year of the movie as an attribute for the corresponding arc, and longer chains can be created just if arcs have years in the increasing order. However, such essentially more correct approach would complicate graph analysis, and the author did not meet such in the literature. Moreover, questions about the existence and content of a possible “flow” between indirectly “tied” movie actors are still open.

6.2 Collaboration network and Erdős numbers

Another popular example is the network of joint publications [54]. Each collaboration between coauthors of particular publication constituting the basis of the built network is correct – each vertex corresponds to a particular author and edge between two vertices denotes mutual publication and, most probably, also real collaborative work. Several joint publications may be represented by separate edges or by one weighted edge where weight is the number of joint publications. If analysts investigate particular author, all is correct until they do not cross the border of distance one where ends collected data. Investigating things beyond this (say at a distance two from a particular vertex) mirrored back to real people needs additional explanation. The special case of collaboration network is attributed graph where “distance” from the famous mathematician Paul Erdős (1913 - 1996) [57] is investigated [32,58]. The network is also mentioned in [12].

“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." [59].

However, what **exactly** means "are linked through the papers" for distances greater than 1, i.e. for persons not being co-authors?

Obviously, it is assumed that real mathematicians (if the corresponding vertices belong to the Erds component of the collaboration graph) **have** finite Erdős numbers like names and surnames - this "property" from the graph is mirrored back to real life ("Once you know your Erds number, you can use it in various ways, such as your license plate number." [58]). We can find "The point is that most mathematicians have Erdős numbers of at most 4 or 5, and – extending the collaboration graph to include co-authorship across all the sciences – most scientists in other fields have Erdős numbers that are comparable or only slightly larger; Albert Einstein's is 2, Enrico Fermi's is 3, Noam Chomsky's and Linus Pauling's are each 4, Francis Crick's and James Watson's are 5 and 6 respectively." [32]. Or "There are five other people with means less than 5. In order of increasing mean, they are Ronald Graham, Andrew Odlyzko, Noga Alon, Larry Shepp, and Frank Harary." [58].

Having lower Erdős number means producing high-quality publications? Is it enough to announce Erdős number as a proof of quality and the author will pass reviewing procedure to get published? Rather not. Similarly to problems with the network of movie actors, since the death of P.Erdős in 1996 today can not exist any tangible ties with him. If lower Erdős number implies higher scientific level (in any reasonable and verifiable way), then it would be equivalent to claiming that each next generation of scientists publishing their papers is of lower scientific level if compared with the previous one (since death of P.Erdős it is impossible to get Erdős number higher than 2, after passing away of all Erdős co-authors there will be impossible to get values higher than 3, and so on). Colleagues noticed an interesting feature justifying that Erdős number can not be a measure of "quality" of a particular scientist. It is possible that Erdős number of a particular author is decreased (scientific "quality" increased) by doing absolutely nothing [60]. This situation is explained graphically in Fig. 5 for the author "X" – it is enough if some author on "X social path" decrease Erdős number by publishing a paper with co-author having a less Erdős number and as a consequence, the number is decreased for a group of connected authors.

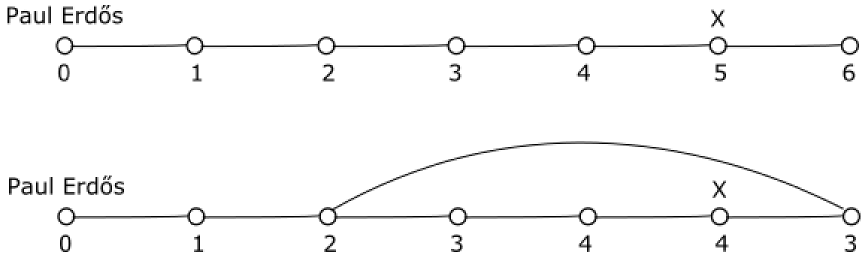

Fig. 5. Decreasing Erdős number of X by doing nothing.

As well assigning numbers starting from P.Erdős suggests an idea that some imaginary “flow” is going from P.Erdős and instead of actual collaboration, we get some “advisory flow” from co-authors with lower Erdős numbers to authors with lower numbers. Therefore Erdős numbers can not be considered to be an accurate measure to reflect collaboration and spreading scientific ideas for all joint publications.

7 Conclusions

Graphs are a powerful tool for the analysis of networks, and usually, concepts and constructions from real networks are identified with graph concepts without reasonable criticism. In some cases, usage of graphs can not be admitted as correct, especially if direct ties represent static facts. Assuming that social networks with intransitive relationships can be modeled in the same way as physical networks together with graph metrics based on the concepts of path and connectivity via indirect ties are root causes of observed problems.

In several cases, social network analysts simply switch from network to graph model without proving that graph-based transformations and calculations are transformable back to the network.

With the rise of machine learning more and more effort must be put on validating of the obtained results to the network. Mechanical transformation of results back to the real life and proceeding by them without reasonable criticism and proven correctness may be dangerous.

These findings could help social network analysts to look more critically at their models as well to reconsider conclusions obtained from a graph based network analysis.

Acknowledgements

This work was supported by Latvian State research program “Next generation information and communication technologies” (NexIT), 2014 - 2017.

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

References

1. Butts, C.T.: Revisiting the foundations of network analysis. Science **325**(5939) (2009) 414–416
2. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press (1994)
3. Hopkins, B., Wilson, R.J.: The Truth about Königsberg. The Colledge Mathematics Journal **35**(3) (5 2004) 198–207
4. de Nooy, W., Mrvar, A., Bategelj, V.: Exploratory Social Network Analysis with Pajek. 2 edn. Cambridge University Press (2012)

5. Diestel, R.: Graph Theory. Graduate Texts in Mathematics. Springer-Verlag Berlin Heidelberg (2017)
6. Seidman, S.B.: Structures induced by collections of subsets: A hypergraph approach. Mathematical Social Sciences (1) (1981) 381–396
7. Scott, J., ed.: Social Networks: Critical Concepts in Sociology. Volume 1. Routledge (2002)
8. Scheidel, W., Meeks, E., Grossner, K., Alvarez, N.: Orbis – the stanford geospatial network model of the roman world `http://orbis.stanford.edu/`.
9. Bothorel, C., Cruz, J.D., Magani, M., Micenková, B.: Clustering attributed graphs: Models, measures and methods. Network Science **3**(3) (2015) 408444
10. Denny, M.: Institute for Social Science Research, University of Massachusetts Amherst, Workshop "Social Network Analysis" (2014) `http://www.mjdenny.com/workshops/SN_Theory_I.pdf`.
11. Borgatti, S.P., Everett, M.G., Johnson, J.C.: Analyzing Social Networks. SAGE publications Ltd. (2013)
12. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393** (1998) 440–442
13. Kleinfeld, J.S.: The Small World Problem. Society **39**(2) (January 2002) 61–66
14. Duchon, P., Hanusse, N., Lebhar, E., Schabanel, N.: Could any graph be turned into a small-world? Theor. Comput. Sci. **355**(1) (April 2006) 96–103
15. Tufte, E.R.: Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire, CT, USA (1997)
16. Robinson, W., Boisjoly, R., Hoeker, D., Young, S.: Representation and Misrepresentation: Tufte and the Morton Thiokol Engineers on the Challenger. Science and Engineering Ethics **8**(1) (2002) 59–81
17. Krebs, V.E.: Social capital: the key to success for the 21st century organization. IHRIM **XII**(5) (2008) 40
18. Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. Physics Review (2003) E 68:036122.
19. Krebs, V.: Social network analysis: An introduction by orgnet, llc. `http://www.orgnet.com/sna.html` (2002)
20. Kleinfeld, J.S.: Could It Be a Big World? (2001) `http://www.judithkleinfeld.com/ar_bigworld.html/`.
21. Akoglu, L., McGlohon, M., Faloutsos, C. In: oddball: Spotting Anomalies in Weighted Graphs. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 410–421
22. Fortunato, S.: Community detection in graphs. Physics Reports **486** (February 2010) 75–174
23. 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.: Passim – an open source software system for managing information in biomedical studies. BMC Bioinformatics **8**(1) (2007) 1–7
24. Opmanis, M., Čerāns, K.: Multilevel data repository for ontological and metamodeling. In: Databases and Information Systems VI-Selected Papers from the Ninth International Baltic Conference, DB&IS. (2010)
25. Čerāns, K., Vīksna, J.: 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 (1996) 389–400
26. Luce, R.D., Perry, A.D.: A method of matrix analysis of group structure. Psychometrika **14**(2) (1949) 95–116

27. Fibich, G.: Diffusion of new products with recovering consumers. https://arxiv.org/abs/1701.01669v2 (2017)
28. Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O.: Gossip: Identifying Central Individuals in a Social Network. Working Papers id:5925, eSocialSciences (June 2014)
29. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery **29**(3) (May 2015) 626–688
30. Bondy, J.A., Murty, U.S.R.: Graph Theory With Applications. Elsevier Science Publishing Co., Inc. (1976)
31. Peay, E.R.: Connectedness in a General Model for Valued Networks. Social Networks (2) (1980) 385–410
32. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press (2010)
33. Borgatti, S.P.: Centrality and network flow. Social Networks **27**(1) (2005) 55 – 71
34. Faloutsos, C., McCurley, K.S., Tomkins, A.: 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 (2004) 118–127
35. Bavelas, A.: Communication patterns in task oriented groups. The Journal of the Acoustical Society of America **22**(6) (1950) 725–730
36. Kilduff, M., Krackhardt, D.: Interpersonal Networks in Organizations. Cognition, Personality, Dynamics, and Culture. Cambridge University Press (2008)
37. Milgram, S.: The Small World Problem. Psychology Today **2** (1967) 60–67
38. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. Journal of the American Statistical Association **97**(460) (2002) 1090–1098
39. Carpenter, D.P., Esterling, K.M., Lazer, D.M.J.: Friends, brokers, and transitivity: Who informs whom in washington politics? Journal of Politics **66**(1) (2004) 224–246
40. Bekmambetov, T., Chevazhevskiy, Y., Jonynas, I., Kiselev, D., Voytinskiy, A.: Movie "Six Degrees of Celebration" (original title – "Yolki"). http://www.imdb.com/title/tt1782568/ (2010)
41. Blackmore, S., Dawkins, R.: The Meme Machine. New ed edn. Oxford University Press (2000)
42. Goffman, W., Newill, V.A.: Communication and Epidemic Processes. Proceedings of the Royal Society of London Series A **298** (May 1967) 316–334
43. Goffman, W.: A mathematical method for analyzing the growth of a scientific discipline. J. ACM **18**(2) (April 1971) 173–185
44. Funk, S., Gilad, E., Watkins, C., Jansen, V.A.A.: The spread of awareness and its impact on epidemic outbreaks. Proceedings of the National Academy of Sciences **106**(16) (2009) 6872–6877
45. Kleinberg, J.: 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 (2000) 163–170
46. Travers, J., Milgram, S.: An Experimental Study of the Small World Problem. Sociometry **32**(4) (December 1969) 425–443
47. James, R., Croft, D.P., Krause, J.: Potential banana skins in animal social network analysis. Behavioral Ecology and Sociobiology **63**(7) (2009) 989–997
48. Hernández, J.M., Mieghem, P.V.: Classification of graph metrics. Technical report, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 CD Delft (November 2011)

49. Park, K., Yilmaz, A.: A social network analysis approach to analyze road networks. In: ASPRS Annual Conference. San Diego, CA. (2010)
50. Park, K.J., Yilmaz, A.: Social network approach to analysis of soccer game. In: Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE (2010) 3935–3938
51. Vihrovs, J., Prūsis, K., Freivalds, K., Ručevskis, P., Krebs, V.: An inverse distance-based potential field function for overlapping point set visualization. In: Proceedings of the 5th International Conference on Information Visualization Theory and Applications (VISIGRAPP 2014). (2014) 29–38
52. Ķikusts, P., Ručevskis, P.: Personal conversation (2013)
53. Needham, C.: Internet movie database. http://www.imdb.com (1998)
54. Borenstein, E.: University of Washington course GS559: Introduction to Statistical and Computational Genomics (Winter 2016), Slides of lecture 15: Biological networks and Dijkstra's algorithm (2016) http://elbo.gs.washington.edu/courses/GS_559_16_wi/slides/15A-Networks_Dijkstra.pdf.
55. Fass, C., Turtle, B., Ginelli, M.: Six Degrees of Kevin Bacon. Plume (1996)
56. Connery, S.: Filmography. http://www.seanconnery.com/filmography/ (2016)
57. Erdős, P.: Wikipedia. https://en.wikipedia.org/wiki/Paul_Erd%C5%91s (2016)
58. Grossman, J.W.: The Erds Number Project (2015) https://oakland.edu/enp/.
59. Barabasi, A., Frangos, J.: Linked: The New Science Of Networks Science Of Networks. Basic Books (2014)
60. Ručevskis, P., Podnieks, K., Kozlovičs, S., Grasmanis, M., Celms, E.: Personal conversation (2016)