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. Only in rare cases results obtained from the graph models are validated against "ground truth" and are directly applicable to objects in the investigated domain. Like extraneous solutions in mathematics, ungrounded mechanistic analogies and incorrect interpretation of indirect ties for intransitive relations as well as usage of "path" concept for social networks may lead to not invertible results having no evidence outside the used graph model. The author investigates unimodal networks with dyadic ties, provides several examples of correct and incorrect applications as well as recover roots of incorrectness.

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

1 Introduction

Together with physical networks like transportation and computer-related networks, also social networks comprising *actors* (humans or human-based structures like companies, parties, and social groups) and *relationships* (ties, interactions) between them are investigated via attributed graphs. Excellent general overview of the history of graph usage in social network analysis is given in [1], while [2] contains 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. Example of such social network is depicted in Figure 1.

We will provide a clear line between *network* as real-world phenomena and its model – *graph*. Term "graph" here is used in the narrow sense of the word exclusively in connection with graph theory and has nothing with things like infographics, charts, and functions. It is assumed that relationships in the network should be verifiable without using any model including graphs.

Such 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." [3] In others network terms are simply given as "synonyms" of graph terms [4]: "Actor:

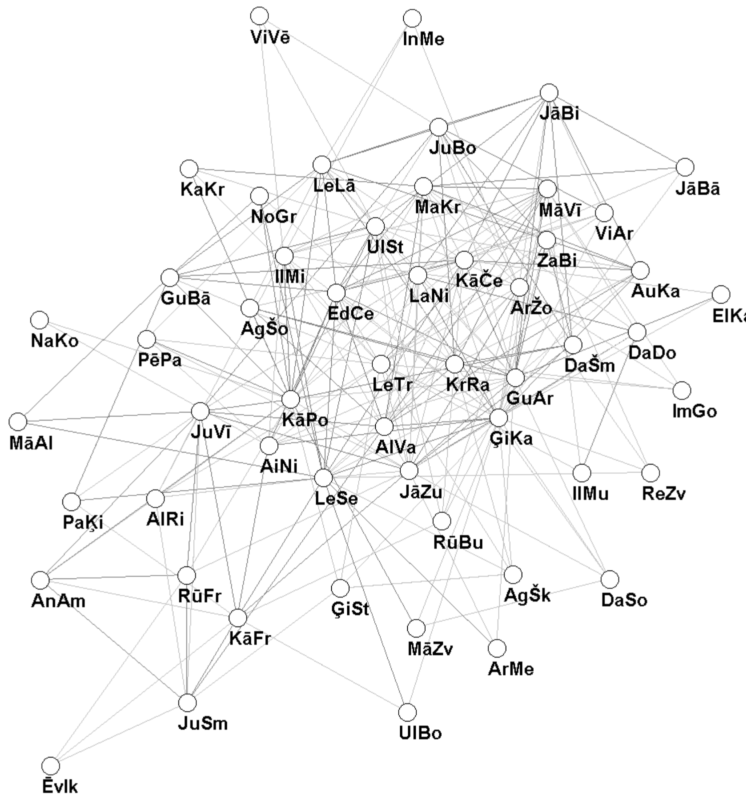

Fig. 1. Attributed graph of the social network.

also called a node or a vertex" [5], "Most often, nodes are individuals, such as individual persons or chimpanzees." [6], "... the propagation of a sexually-transmitted disease that spreads along the edges of a graph." [7].

Despite the fact that such interviewing justifies naturality of graph concepts for network analysis, it puts 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 not clearly subdividing whole process in a separate steps. If network and graph terms are used interchangeably, this gives 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 **always** is 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 there is chosen different network model instead graphs.

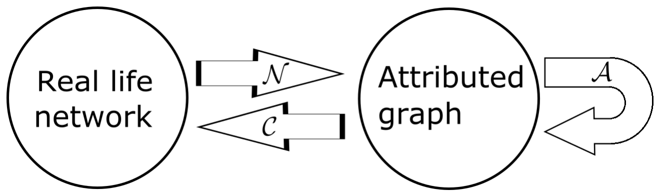

Fig. 2. The process of network analysis using graphs: $\mathcal{N}$ – obtaining attributed graph, $\mathcal{A}$ – performing analysis, $\mathcal{C}$ – applying analysis results

In this paper, we will assume that 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 in the literature there are mentioned several sources of incorrectness of these steps. For example, speaking about social networking services in [8]: "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" [4] (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." [9]

Attention to the correctness of this step in the social network analysis literature is surprisingly low. Just a few authors [10,11] emphasize necessary to validate results obtained from graphs with respect to the original network. The value of network structure investigation separately also is 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." [12]

Similar critical thoughts aimed at inappropriate usage of numbers in general we can found in [13]: "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."

In this paper, we will demonstrate that concept of "path" as a chain of consecutive ties or "connectivity" which is natural for graphs and have good analogs in substantial networks is 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 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 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 reader is familiar with graph concepts [2,9,14], but would like to briefly remind 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 order of objects in pairs is important) or *edges* (if order is not important).

A particular graph by definition is **static** structure - V and E are fixed and "analysis of graph" means analyzing these two sets. Straightforward outcome from this fact is that graph models of dynamic networks can be just snapshots at some time moment or describe underlying static structure.

The graph itself doesn't contain "historical" information how sets V and E are created and why these sets contain exactly these elements. "Meaning" of V and E is out of scope from the viewpoint of the defined graph. Therefore, if there is intent to apply results obtained from the graphs to the initial network, meaning should be somehow kept beside just bare graph sufficient for graph-based analysis. The simplest form is adding attributes 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 described domain and labels are too weak for a proper "decoding". Let's investigate one more example.

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

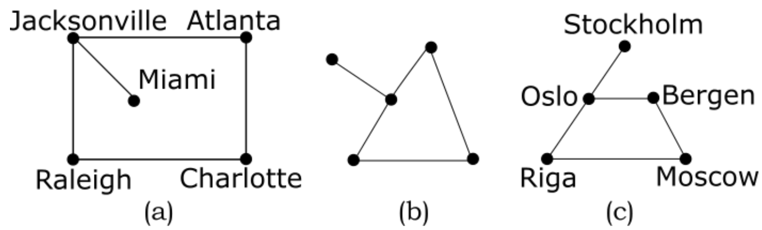

Fig. 3. Isomorphic graphs.

Assuming that (a) and (c) depicts 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?
- Is 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) relationship means "is connected by railroad", it becomes possible to give partial answer to the first question: "In (a) small fragment of USA railroad network is depicted" as well give negative answers to the last two questions since Riga and Oslo is not connected by railroad and therefore relationship in (c) obviously differs from (a).

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

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

3 From network to graph

Let's investigate simple example how graph can be obtained from a particular network. Let's try to describe 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 investigation of collaborative work of movie actors and we will be interested in a relationships "Actors X and Y performed together in the same episode". What is the most appropriate way how to build the corresponding graph?

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

However, this is not the only way. If we use multi-layer graphs [15] and focus on a separate movies, we can create separate graph (or graph "on a separate sheet") for each movie, like in the Fig. 4(b). In the movie M_1 performed A, B, C, E and F, while in the M_2 performed B, C, D, E and F. As a result, there can be several vertices representing the same person in a different movies.

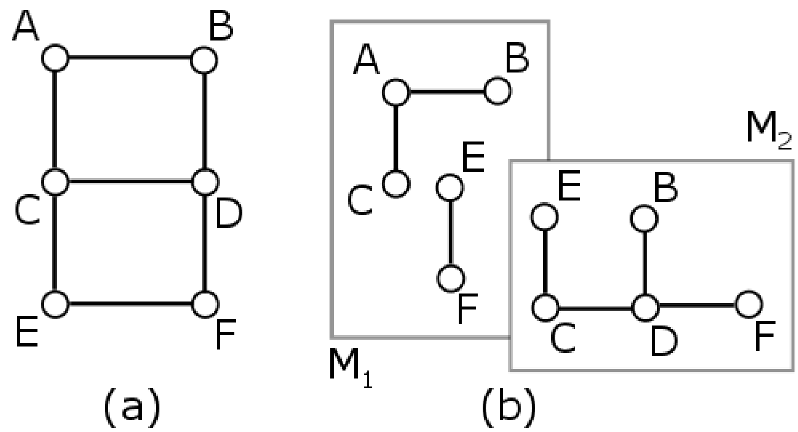

Fig. 4. Graphs obtained from the same network – (a) neglecting particular movie information, (b) multi-layer graph with separate graph for each movie

It should be pointed out that facts obtained from 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 its simplicity "all in one" way of modeling is preferred by network analysts, while other possible approaches are not investigated. However, conclusions obtained from graphs can essentially differ depending on the chosen approach. In our example question "Does the actors X and Y ever performed in the same episode?" can be answered from both versions while "Does the actors X and Y ever performed in the same movie?" cannot be answered from 4(a) if there is no edge between particular vertices. So, answer is "yes" for D and E while "no" for their structural analog D and A.

Chosen graph model highly depends on research purpose. For example, if the intention is to investigate pairwise collaboration for a particular actor expressive characteristic of each vertex (an *ego*) is obtained by investigating its induced 1-step sub-graph (referred as *egonet*) [16]. In the given example egonet with ego B can be better explored directly in the "all (collaborations) in one" graph (Fig. 4 (a)). To obtain number of different B partners in some episode, we should only calculate degree of vertex B (2). Collection of the same information for the multi-layer graph needs some preprocessing like creation of virtual vertex B' where all appearances of B are collected together.

4 Direct and indirect ties

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

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

Fig. 5. Graph of three friendships

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 should take a look at the corresponding attributed graph of friendships, find 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. Building the graph 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 graph – "yes/no" answer of all involved persons X to the question "Is Y your friend?" should conform to the previously gathered information.

But what we can say about an indirect relationships between pair of persons not tied directly, like A and D? Definitely, they are not friends (absence of an edge between corresponding vertices). Are they familiar? Maybe yes (but then they are not friends) and maybe not – relationship "is familiar" was not described in the initial set of facts for persons not being friends and therefore is not presented in the graph regardless way of coding. As a consequence, it is not possible to decide which option takes place without additional information about relationships besides friendship in the observed network.

Since network of friends is a popular standard example and several authors speak about "transitivity of friendship" in terms "it is a tendency for friends of friends to be friends" [5] or "the enemy of my enemy is my friend" [6, p.22]. In real examples "a friend of a friend is friend" may be "with high probability" [17] but far from taking place always.

There is always a possibility to introduce artificial indirect relationships like "secondary friends"(i.e. there is common friend for both), but such relationships are not simply observable in the initial network without seeing graph. For example, for a randomly chosen person it should be hard to correctly answer question "Is Y you secondary friend?" for all persons besides friends in the network.

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

5 Path concept

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

Definition 2. *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. Due to 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 steps performed.

Definition 3. *Length of a path* is number of edges in this path.

Also, we can introduce term "connectivity".

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

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

Definition 6. *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 7. *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 [18].

Connectivity in graphs as well as usage of terms "walk", "trail", "path" [19, p.12] is so intrinsic that social network analysts neglect the necessity to define corresponding constructs in the investigated domain and takes for the granted meaningful existence of them also there. In [20] 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.

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 ($\mathcal{S}_N$).

A typical representative of the social network model is a graph where it is possible to find a connected component which is not a clique and therefore belongs to $\mathcal{S}_N$. Representatives of these classes are depicted in Fig. 6.

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." [4]. Similarly, "Those nodes

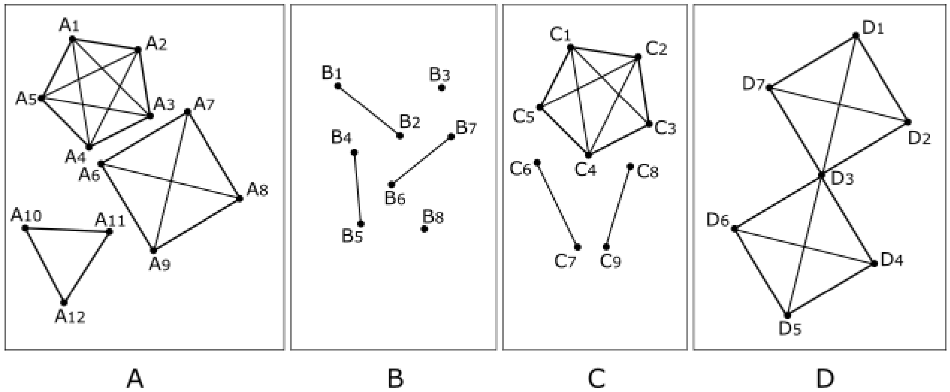

Fig. 6. Representatives of $\mathcal{S}_C$ (A, B) and $\mathcal{S}_N$ (C, D).

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." [16]. And, "Obviously, social networks are neither complete not one-dimensional." [21].

If there are separate connected components, they could be investigated separately [22]. 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 look for anomalies in the graphs [23].

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

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

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

Proof. Since $g \in \mathcal{S}_\mathcal{C}$, all connected components $c \subseteq g$ are cliques, then 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 clique is connected with all other vertices in clique. Therefore transitivity requirement is fulfilled: $v_i E v_j$ and $v_j E v_k$ imply $v_i E v_k$. $\square$

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

Proof. Since $g \in \mathcal{S}_N$, 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 – otherwise 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 $v_i E v_k$. $\square$

Proposition 3. Relationship P over the set of all $g \in \mathcal{S}_N$ 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: $v_x P v_z$. $\square$

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

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

Propositions show that in the case of $\mathcal{S}_N$ there is the essential difference between direct and indirect ties (or paths having length 1 and greater than 1) – direct ties **cannot be simply considered** as a special case of longer paths or paths automatically having 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 secret that ties in this network are defined as “there exists railroad connection between cities **or** there is the same number of letters in the names of cities written in English”. The provided graph is formally correct while of low value for investigating indirect ties in the initial network of cities. It is not obvious that there is any valuable relationship between not connected cities defined for any pair of them. Until defining such relationship (analog to P in the theoretical model), there is no sense to talk about anything based on the path concept.

Only having meaningful path explanation it is worth to calculate distances between vertices, seek for shortest paths between pairs of vertices and calculate an overwhelming number of different graph *metrics* to analyze graph properties.

7 Roots of an incorrect application of graphs

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. However, we can clearly see difference between static structure (road, wires) and dynamic processes which use this structure (someone walking, electric current passing).

A usual way to explain social networks is providing an analogy with the *static* structure of some physical network and further exploit analogy of *dynamics* on an intuitive basis. Road or pipeline networks as well as electric circuits [24] are a few such analogs.

In [6, 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.” [25]. “Information flows” are also mentioned in [8]: “Employees who are included in key information flows and communities of knowledge are more dedicated and have a much higher rate of retention.” In [26] “attitude influencing” and “emotional support” are mixed together with “e-mail broadcast” and “mitotic reproduction”.

Semantics of terms “walk”, “trail”, “path” assumes dynamics – 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. Network modeling by graphs implies “possibility to travel” via edges or chain of consecutive edges without limitations. Like investigating description of the network – famous bridges of Königsberg by Leonhard Euler (and considered being the first paper in graph theory) [27] it is assumed that there are no limits to walk using any of the available pathways.

All topological metrics of distance class used also for exploring social networks (like diameter, betweenness centrality, closeness centrality and eigenvector centrality) are based on concept “path in a graph” [28].

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. Author insists that 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, surfing via 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.

If for the physical networks dynamic processes are justified (like electric current can pass 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 [29]. 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 [30], [1, p.76], [31, 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." [32] The similar thoughts (when speaking about graph diameter) you can find in [5]: "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 [11]: "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!

8 Transmitting messages over networks

As a good mental exercise investigation of the relation "sends messages to" already described in [18] for two networks may be used: 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 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" [33], 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 [34]. Most probably we will get "Chinese whispers" [35] 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 [36,37]. As pointed out in [38]: "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 several observations.

First, any message can survive 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." [22]). In [39, p.206] 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 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)." [40].

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." [22]. 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." [40]. As well information can be carried in disagreement with physical laws in their mechanic analogs: "Flow betweenness counts all paths that carry information when a maximum flow is pumped between each pair of vertices. In many networks, however, neither of these cases is realistic. Both count only a small subset of possible paths between vertices, and both assume some kind of optimality in information transmission (shortest paths or maximum flow)" [41]

Similar doubts author can find only in the papers describing a few known **real** experiments with the usage of social ties [32,42]. 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 [43]. Also in later experiment [12] 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 [44].

An excellent conclusion is given in [41]: "And even in a case such as the famous small-world experiment of Milgram [32] and Travers and Milgram [42], or its modern-day equivalent [12], 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 natural and quite obvious meaning for indirect ties. 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 with directed relationships. It should be pointed out, that while these networks are representing society, 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 [45], where road map of Roman Empire can be investigated looking for shortest, fastest or cheapest routes. Various interesting results can be obtained by calculations 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 provided on the graph are backward compatible with the initial network.

9.2 Consanguinity

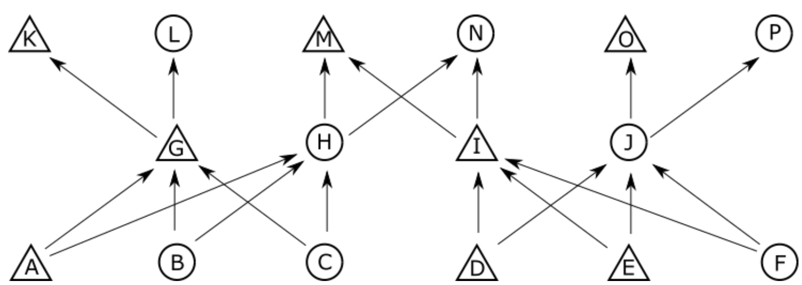

Fig. 7. Graph of consanguinity network.

The next example is a graph of consanguinity where depicted network consists of people tied with “is a child of” relationship. Consanguinity is defined as “being related to someone by birth” or “having a common ancestor”. Example of such graph is given in the 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 perspective and it is possible to determine *degree of kinship* as a length of a specific path first going upwards (from children to parent) and then downwards (from parent to children). Any of these parts may be absent, but cannot be interchanged. With this restriction distance (or degree) between people in this graph is measured in a way which completely corresponds to the Definition 3. For example, from the A perspective, degree 1 have parents G and H, degree 2 – grandparents K, L, M and N, and sisters B and C, degree 3 – uncle I, and degree 4 – cousins D, E and F.

It should be pointed out that calculating length of an arbitrary path without described restriction we can conclude that two people (e.g., K and P) are connected what may be completely wrong in terms of the initial network if there is no common ancestor.

9.3 Movie actor collaboration

The popular example used in social network analysis is movie actor collaboration network which is built using data from the Internet Movie Database (IMDb) [46,47]. This undirected graph is built by 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" [48,49] is based on these data.

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 [50]. The corresponding attributed graph is depicted in Fig. 8 a). Since W.Lawson and C.Zeta-Jones never performed in the same movie, "distance" between W.Lawson and C.Zeta-Jones according to the graph by definition is 2.

Fig. 8. Actor collaboration.

Traditionally, finite distances (in opposite to infinite) is the sign that 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 non-existing W.Lawson.

Also, existence and content of a possible "flow" between indirectly "connected" actors has not proven also for persons being alive.

9.4 Collaboration network and Erdős numbers

Another popular example is the network of joint publications [47]. Each collaboration between coauthors of particular publication constituting the basis of the built network is correct – each vertex corresponds to a particular author, an edge between two vertices denotes mutual publication and, most probably, also real collaborative work. The special case of collaboration network is attributed graph where "distance" from the famous mathematician Paul Erdős (1913 - 1996) [51] is investigated [25,52]. "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." [53]. The network is also mentioned in [7,54]. American Mathematical Society offers the free online tool to determine Erdős number of any particular author [55].

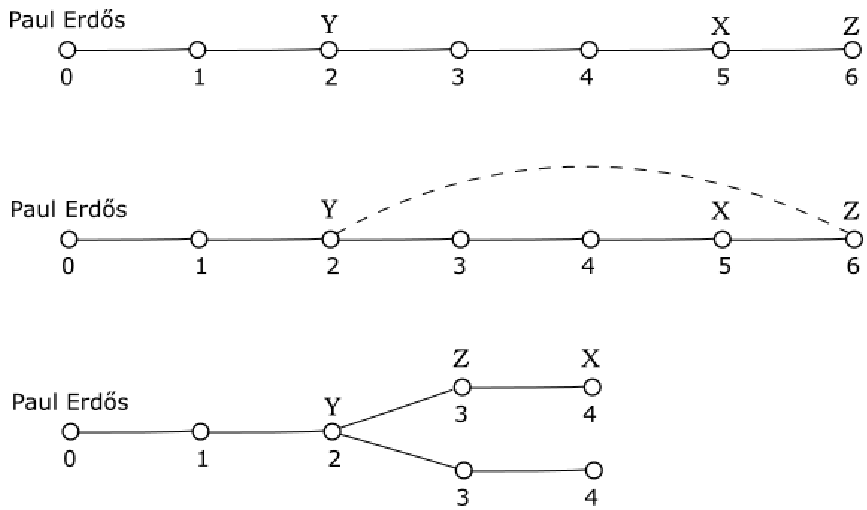

Fig. 9. Decreasing Erdős number of X without direct involvement of X. a) initial state, b) new $Y - Z$ publication, c) updated state

In several sources is given the impression that less Erdős number is somehow related to a higher scientific value of a particular author. However, what **exactly** means "are linked through the papers" for distances greater than 1, i.e. for persons not being co-authors? Having lower Erdős number means producing high-quality publications "by default" or it is enough to announce Erdős number as a proof of quality and the author will pass reviewing procedure to get published? Rather not. At least author's personal experience shows that the same Erdős number may have authors with uncomparable scientific capacity.

An interesting justification that Erdős number cannot be a stable measure of "quality" of a particular scientist is the following [56]: It is possible to decrease Erdős number of a particular author X without involvement of X himself – it is enough if some author Y on "X social path to Erdős" publish a paper with X co-author Z and as a consequence decrease also Erdős number for X (see Fig. 9).

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

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

9.5 Disciplinarity of publications

There may be the attempt to decide disciplinarity of publications from the collaboration network [57]. 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 having three pairwise connected publications [58,59,60] with content not related to the scientific interests of the third.

10 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 cannot 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.

If it is intended to go beyond ego and use graph metrics based on paths in graphs, transitivity check (possibility to interpret indirect ties and prove their transitivity) of ties in the observed network is crucial. If object-based graph model is used, the simple check may be done by switching to fact-based model. Without any such proof, graph metrics based on path concept should not be used and conclusions based on indirect ties should not be made.

As well usage of popular social graph examples like collaboration graphs should be revisited from the viewpoint of internal meaning and usefulness of the obtained numerical values.

With the rise of machine learning more and more effort should 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 is unacceptable.

Acknowledgements

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

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

References

1. Scott, J., ed.: Social Networks: Critical Concepts in Sociology. Volume 1. Routledge (2002)
2. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press (1994)

3. Robins, G.L.: Doing Social Network Research: Network-based Research Design for Social Scientists. 1 edn. SAGE publications Ltd. (2015)
4. Bothorel, C., Cruz, J.D., Magani, M., Micenková, B.: Clustering attributed graphs: Models, measures and methods. Network Science **3**(3) (2015) 408444
5. 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`.
6. Borgatti, S.P., Everett, M.G., Johnson, J.C.: Analyzing Social Networks. SAGE publications Ltd. (2013)
7. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393** (1998) 440–442
8. Krebs, V.E.: Social capital: the key to success for the 21st century organization. IHRIM **XII**(5) (2008) 40
9. de Nooy, W., Mrvar, A., Bategelj, V.: Exploratory Social Network Analysis with Pajek. 2 edn. Cambridge University Press (2012)
10. Krebs, V.: Social network analysis: An introduction by orgnet, llc. `http://www.orgnet.com/sna.html` (2002)
11. Kleinfeld, J.S.: Could It Be a Big World? (2001) `http://www.judithkleinfeld.com/ar_bigworld.html/`.
12. Dodds, P.S., Muhamad, R., Watts, D.J.: An experimental study of search in global social networks. Science **301**(5634) (2003) 827–829
13. Edmonds, D.B.: Against the inappropriate use of numerical representation in social simulation. (2004)
14. Diestel, R.: Graph Theory. Graduate Texts in Mathematics. Springer-Verlag Berlin Heidelberg (2017)
15. Kim, J., Lee, J.G.: Community detection in multi-layer graphs: A survey. SIGMOD Rec. **44**(3) (December 2015) 37–48
16. Akoglu, L., McGlohon, M., Faloutsos, C. In: oddball: Spotting Anomalies in Weighted Graphs. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 410–421
17. 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
18. Luce, R.D., Perry, A.D.: A method of matrix analysis of group structure. Psychometrika **14**(2) (1949) 95–116
19. Bondy, J.A., Murty, U.S.R.: Graph Theory With Applications. Elsevier Science Publishing Co., Inc. (1976)
20. Peay, E.R.: Connectedness in a General Model for Valued Networks. Social Networks (2) (1980) 385–410
21. Fibich, G.: Diffusion of new products with recovering consumers. `https://arxiv.org/abs/1701.01669v2` (2017)
22. 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)
23. 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
24. Bozzo, E., Franceschet, M.: Resistance distance, closeness, and betweenness. Social Networks **35**(3) (2013) 460 – 469
25. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press (2010)
26. Borgatti, S.P.: Centrality and network flow. Social Networks **27**(1) (2005) 55 – 71

27. Hopkins, B., Wilson, R.J.: The Truth about Königsberg. The Colledge Mathematics Journal **35**(3) (5 2004) 198–207
28. 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)
29. 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
30. Bavelas, A.: Communication patterns in task oriented groups. The Journal of the Acoustical Society of America **22**(6) (1950) 725–730
31. Kilduff, M., Krackhardt, D.: Interpersonal Networks in Organizations. Cognition, Personality, Dynamics, and Culture. Cambridge University Press (2008)
32. Milgram, S.: The Small World Problem. Psychology Today **2** (1967) 60–67
33. 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
34. 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)
35. Blackmore, S., Dawkins, R.: The Meme Machine. New ed edn. Oxford University Press (2000)
36. Goffman, W., Newill, V.A.: Communication and Epidemic Processes. Proceedings of the Royal Society of London Series A **298** (May 1967) 316–334
37. Goffman, W.: A mathematical method for analyzing the growth of a scientific discipline. J. ACM **18**(2) (April 1971) 173–185
38. 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
39. Kadushin, C.: Understanding Social Networks: Theories, Concepts, and Findings. 1 edn. Oxford University Press (2012)
40. 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
41. Newman, M.J.: A measure of betweenness centrality based on random walks. Social Networks **27**(1) (2005) 39 – 54
42. Travers, J., Milgram, S.: An Experimental Study of the Small World Problem. Sociometry **32**(4) (December 1969) 425–443
43. Kleinfeld, J.S.: The Small World Problem. Society **39**(2) (January 2002) 61–66
44. Schnettler, S.: A small world on feet of clay? a comparison of empirical small-world studies against best-practice criteria. Social Networks **31**(3) (2009) 179 – 189
45. Scheidel, W., Meeks, E., Grossner, K., Alvarez, N.: Orbis – the stanford geospatial network model of the roman world http://orbis.stanford.edu/.
46. Needham, C.: Internet movie database. http://www.imdb.com (1998)
47. 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.
48. Fass, C., Turtle, B., Ginelli, M.: Six Degrees of Kevin Bacon. Plume (1996)
49. Collins, J.J., Chow, C.C.: It's a small world. Nature **393** (Jun 1998) 409

50. Connery, S.: Filmography. http://www.seanconnery.com/filmography/ (2016)
51. Erdős, P.: Wikipedia. https://en.wikipedia.org/wiki/Paul_Erd%C5%91s (2016)
52. Grossman, J.W.: The Erds Number Project (2015) https://oakland.edu/enp/.
53. Barabasi, A., Frangos, J.: Linked: The New Science Of Networks Science Of Networks. Basic Books (2014)
54. Pelikán, J.: Paul Erdös (1913-1996). Mathematics Competitions **9**(2) (1996) 15 – 20
55. American Mathematical Society: MathSciNet - FreeTools. https://mathscinet.ams.org/mathscinet/freeTools.html?version=2 (2018)
56. Ručevskis, P., Podnieks, K., Kozlovičs, S., Grasmanis, M., Celms, E.: Personal conversation (2016)
57. Fortunato, S.: Community detection in graphs. Physics Reports **486** (February 2010) 75–174
58. 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
59. Opmanis, M., Čerāns, K.: Multilevel data repository for ontological and meta-modeling. In: Databases and Information Systems VI-Selected Papers from the Ninth International Baltic Conference, DB&IS. (2010)
60. 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