Towards Causal Explanations of Community Detection in Networks

Georgia Baltsou[1], Anastasios Gounaris[1], Apostolos N. Papadopoulos[1], and Konstantinos Tsichlas[2]

[1] School of Informatics, Aristotle University of Thessaloniki, Greece
{georgipm,gounaria,papadopo}@csd.auth.gr
[2] Department of Computer Engineering and Informatics, University of Patras, Greece
ktsichlas@ceid.upatras.gr

Abstract. Community detection is a significant research problem in Network Science since it identifies groups of nodes that may have certain functional importance - termed communities. Our goal is to further study this problem from a different perspective related to the questions of the cause of belongingness to a community. To this end, we apply the framework of causality and responsibility developed by Halpern and Pearl [11]. We provide an algorithm-semi-agnostic framework for computing causes and responsibility of belongingness to a community. To the best of the authors' knowledge, this is the first work that examines causality in community detection. Furthermore, the proposed framework is easily adaptable to be also used in other network processing operations apart from community detection.

Keywords: causality, network analysis, community detection, algorithms

1 Introduction

Imagine someone participating in a social network. Due to an analytics engine that the social network offers for its users, she finds out that she is unintentionally part of a community and asks what are the reasons for her belongingness to this community. She would also wish to become a member of another community - always in the context of the community detection algorithm offered by the analytics engine - and asks what new relations she would have to set up in order to become a member. Note that in this example, one's membership in a community is not explicit but implicit through the social network analytics engine, which affects many aspects of the user's belongingness to the social network (e.g., recommendation of new friends, selection of ads to show, etc.) and thus it is of high importance to the user. In particular, we ask:

1. *What causes the fact that a node u belongs to a community C?* Which are the edges that are responsible for $u \in C$? Can we rank these edges based on the degree of their responsibility for $u \in C$?
2. *What causes the fact that a node u does not belong to a community C?* Which are the new edges that would allow u to become a member of C?

Networks are used to represent data in almost any field, such as transportation systems [7], biological systems [22], and social groups [20], just to name a few. In such networks, certain groups of nodes with particular importance arise, which form the so called communities. The dominant definition of community is a group of nodes that are more densely connected internally than externally [25]. In real-world networks, communities are of major importance since they are related to functional units [24,3]. Communities have also topological properties that are different from those of the network as a whole.

In this work, we formulate such questions related to community detection by using the *structural-model approach* introduced by Halpern and Pearl [11,12]. In particular, we focus on the community detection problem and we define different sets of causes with different degrees of importance with respect to the question at hand. This importance is captured by a measure of *responsibility* [4] for each cause, thus allowing for a ranking of the causes. Moreover, this structural-model approach has the side-effect that other communities may change as a result of a question for a particular node. We introduce a measure for these changes to quantify how the interventions implied by the cause alter the community structure of the network. To the best of our knowledge, we are the first to look at the community detection problem on networks through the lens of causal explanations. In fact, it seems that this is the first time that such a viewpoint is adopted with respect to general network analysis problems.

Related Work. Community detection in general has been a very active field during the last years. There is a plethora of algorithms aiming at finding the best quality communities in networks, based on different evaluation metrics. Those works include both disjoint or overlapping community detection algorithms. For some detailed surveys on the field we refer to [8,14]. However, there is no work on combining causal explanations and community detection. The structural-model approach introduced by Halpern and Pearl [11,12] has been applied mainly to database queries. Meliou et al. [18] transferred these notions to databases. This approach is related to data provenance, lineages and view updates (e.g., deletion propagation) [19]. Inspired by this approach others have applied this structural-model approach to reverse-skyline queries [10], to probabilistic nearest neighbor queries [16], and so on. One more network-related problem where this model has been applied concerns the ranking of propagation history in information diffusion in social networks [27].

Contributions and Roadmap. This work focuses on the application of causality in the community detection problem. Examining causality in community detection in networks is novel in its own right. We suggest a general framework that can be used to find causal relations about the belongingness of nodes to communities. An interesting aspect is that the proposed framework is easily adaptable to other network processing operations apart from community detection. Apart from transferring the concepts in [11,12], we also introduce the concept of discrepancy, as a measure of network changes which occur after specific actions.

The rest of the work is organized as follows. In Section 2 we discuss how the causal model of Halpern and Pearl applies to community detection while in Section 3 we provide a general framework that can compute causal explanations and related metrics. In Section 4 we discuss additional issues and future extensions of the proposed framework.

2 The Causal Model for Community Detection

Here we study the proposed causal model and introduce some fundamental concepts.

2.1 Preliminaries

Initially, we restrict our setting to a simple undirected, unweighted network $G = (V, E)$, which is composed of a node set $V = \{1, ..., n\}$ with $n = |V|$ nodes and an edge set $E \subseteq |V| \cdot |V - 1|$ with $m = |E|$ edges; we discuss extensions to more generic graphs in Section 4. Let $G[S]$ represent the induced sub-graph of the node set S, $S \subseteq V$. The adjacent nodes of a node u, i.e., the nodes connected to u with an edge, are its neighbours: $N(u) = \{u|\{u, v\} \in E\}$. The degree of u is $deg(u) := |N(u)|$, i.e., the number of u's neighbours.

Modularity [21] is a widely used objective function to measure the quality of a network's decomposition into communities and is defined as:

$$Q = \sum_{C=1}^{j} \left[\frac{m_C}{m} - \left(\frac{deg^C}{2m} \right)^2 \right]$$

where j is the number of communities in the network, m_C is the number of intra-community edges of C and deg^C is the sum of degrees of all nodes in C.

2.2 The Proposed Approach

The definition of causality is based on the work of Halpern and Pearl [11]. Based on their definition of actual causality we identify and analyze three main concepts within our context: *i*) endogenous and exogenous pairs of nodes, *ii*) contingency sets and *iii*) responsibility.

In general, the fact that a node, henceforth termed as the *query node*, belongs to a community is mainly determined by its incident edges. However, the non-incident edges affect the composition of query's node community and as a result they can affect indirectly the node's belongingness to it. Similar arguments hold for the non-belongingness of a node to a community.

In a nutshell, we try to identify the existing edges that result in the user's v belongingness to a community. Similarly, we try to pinpoint the non-existent incident edges of v that could put v in a new community. To this end, all possible pairs of nodes in $|V| \cdot |V - 1|$ in the network can be partitioned into *exogenous* and *endogenous* ones[3]. Exogenous pairs of nodes $E_x \subseteq |V| \cdot |V - 1|$ are not considered

[3] The endogenous and exogenous sets differ for different nodes. Also, the endogenous set typically comprises the incident edges connecting this node to its neighbors. Finally, allowing self-loops is not an issue, since if they are irrelevant to the setting, we can simply make them exogenous.

to have a causal effect on the (non-)belongingness of node v to a community and endogenous $E_e \subseteq |V| \cdot |V-1|$ are the pairs of nodes that can in principle infer such causal implications. Note that $E_x \cup E_e = |V| \cdot |V-1|$ and $E_x \cap E_e = \varnothing$.

To check if an edge e is a cause for the (non-)belongingness of a node v to a particular community C, we have to find a set of endogenous pairs of nodes whose edge removal/addition will allow e to immediately affect the belongingness of v to the community C. These sets are called *contingencies*. Note that the contingency set does not alone change the community of v but it is required in order to unlock the causal effect of edge e on the belongingness of v. The contingency set must be minimal, in a manner that removing any edge from it, will dampen the causal effect of e to the belongingness of v to its community, so, no redundancy is allowed.

In a sense, all incident edges (and possibly additional ones) of v affect its belongingness to the community (either in positive or negative manner). Thus, we need a ranking function that will allow us to reason about the most important causes for v (non-)participating in the community. *Responsibility* [4] measures the degree of causality of an edge e for a node v as a function of the smallest contingency set.

In the following, we provide a definition of causality tailored to the problem of why a node belongs to a particular community.

Definition 1. *Let $e \in E_e$ be the edge connecting an endogenous pair of nodes and let v belong to community C.*

- *e is called a counterfactual cause for $v \in C$ if for the network $G = (V, E)$ it holds that $v \in C$ while for $G' = (V, E - \{e\})$ it holds that $v \notin C$.*
- *e is called an actual cause for $v \in C$ if there exists a set of edges $\Gamma \subseteq E_e$ called a contingency for e such that e is a counterfactual cause for $v \in C$ in the network $G' = (V, E - \Gamma)$.*

Next, we provide a definition of causality tailored to the problem of explaining why node v is not a member of a community C'.

Definition 2. *Let $e \in E_e/E$ a non-existent edge and C' a community that does not contain v.*

- *e is called a counterfactual cause for $v \notin C'$ if for the network $G = (V, E)$ it holds that $v \notin C$ while for $G' = (V, E + \{e\})$ it holds that $v \in C$.*
- *e is called an actual cause for $v \notin C$ if there exists a set of edges $\Gamma \subseteq E_e$ such that $\Gamma \cap E = \varnothing$ called a contingency for e such that e is a counterfactual cause for $v \notin C'$ in the network $G' = (V, E + \Gamma)$.*

Finally, based on [4] we provide a measure of the degree of causality, thus providing a ranking function for the various causes.

Definition 3. *Let v be the query node with respect to a community C in the network $G = (V, E)$ and let the set of edges e be a cause. The responsibility of e for v participating or not in C is:*

$$\rho_e = \frac{1}{1 + min_\Gamma |\Gamma|}$$

for all contingency sets Γ for e, where $|\Gamma|$ is the size of the set Γ.

The domain of ρ_e is in $(0, 1]$. If the contingency set is $\varnothing$, then the responsibility is 1, otherwise, the larger the contingency set the less the responsibility. In this way, we capture the degree of interventions needed (the set Γ) to uncover the causal implication of e on v with respect to community C.

At this point we need to discuss a distinguishing feature in the introduction of causality in community detection. The counterfactual interventions suggested by the contingency set and the cause may as well change other communities. This may seem as an undesirable side-effect of our definition that we may choose to ignore, as the question of the causes for the belongingness or non-belongingness of v to community C is related to v alone, and so potential changes to other communities are of no interest to v. Since these causes are counterfactual, in fact no change happens if they are simply used for the purpose of briefing v.

However, if we consider v to be an agent whose purpose is to find out what actions should be taken in order to achieve her removal from C (in the case she asks of the causes of her belongingness to C) or her addition to C' (in the case she asks of the causes of her non-belongingness to C') then this side-effect becomes important. In this case, the causes and their corresponding contingency sets can be considered as a suggested set of actions so that v achieves her goal. Apparently, the endogenous set must be defined so that v can alter the corresponding edges. Going into more depth, one will confront various issues like the identity problem that comes up in community detection in temporal networks [23]; that is, after the intervention, what happens if C has changed so much that cannot be considered as C anymore? We avoid such issues by introducing a measure of such changes, called *discrepancy.*

Definition 4. *Let v be a query node with respect to community C in the network $G = (V, E)$. Let V_c be the set of nodes, excluding v, that change community as a result of the intervention implied by the cause e and its corresponding contingency set E_s. Then, the discrepancy $\gamma(v, e, E_s)$ of v with respect to e and E_s is defined as:*

$$\gamma(v, e, E_s) = \frac{|V_c|}{|V| - 1}$$

The domain of the discrepancy is $[0, 1]$. If it is zero then no node changes and thus $|V_c| = 0$. If all nodes change then $|V_c| = |V| - 1$ and thus discrepancy is equal to 1.

Finally, we make three assumptions for efficiency and effectiveness purposes. These assumptions mostly affect the algorithmic aspects discussed in the next section, but are given here as part of the core proposal. The first assumption that was already implied in the discussion of endogenous pairs of nodes, concerns the edges that constitute causes of the (non-)belongingness of node v to a community C. This is the *Locality Assumption.*

Assumption 1. *The more distant two nodes the less they influence each other.*

This assumption allows us to focus on possible causes around node v and in the involved communities. Pairs of nodes whose corresponding edges are considered far from v and do not belong to the involved communities are not considered as endogenous. In community detection in unweighted networks, this assumption is part of its very definition, since the belongingness of node v to a community is guided mainly by its incident edges. Such an assumption is widely used in network analysis [2,5,6], e.g., in social networks is known as the *Friedkin's postulate* [9].

We also make an assumption concerning the size of communities, called henceforth the *Size Assumption.* This assumption allows us to bound the number of endogenous pairs of nodes introduced by the involved communities.

Assumption 2. *Communities are polynomially smaller than the size of network.*

Usually, communities tend to be smaller than the size of the network. As discussed in [8], after systematic analysis by the authors of [15], communities in many large networks, including traditional and online social networks, technological, information networks and web graphs, are fairly small in size. It is also believed that the communities in biological networks are relatively small i.e., $3 - 150$ nodes [26,28]. We capture this phenomenon by assuming that the size of communities is $O(n^{\epsilon})$ for some small constant $\epsilon < 1$.

Finally, for efficiency reasons, we assume that both the number of causes and the size of the Γ set are small. We call this assumption the *Bound Assumption.*

Assumption 3. *The number of causes and the size of contingency sets are bounded by a small constant.*

This assumption is important because it limits the available options for causes and the contingency set. If the number of causes was large, then the information on the causal relations would be minuscule. Besides, if the size of the contingency sets were large, that would lead to a very low value of responsibility, meaning that the effect of the actual cause is minuscule.

Example. We discuss a simple example to provide a foothold to move to the framework description in the next section. n Figure 1, the friendships between members of the Zachary Karate club [29] are shown. The Louvain method [1] has been used to partition the network into 4 communities. Note that in reality the *Green* and *Orange* communities are the one group after the division while the *Blue* and *Purple* communities correspond to the other group. The modularity Q of this network decomposition is 0.417.

Let us first look at node 10. What is the cause for $10 \in$ *Blue*? Removing edge $(10, 34)$ apparently leads to 10 not belonging anymore in *Blue* but in *Green* and thus edge $(10, 34)$ is a counterfactual cause. This is the case with modularity $Q = 0.427$ and no other node changes community, which means that discrepancy $\gamma = 0$ while responsibility $\rho = 1$, since the contingency set is empty. The Locality Assumption 1 was used since the endogenous pairs of nodes were assumed to be only the neighbours of node 10. In case we extend the set of endogenous pairs to

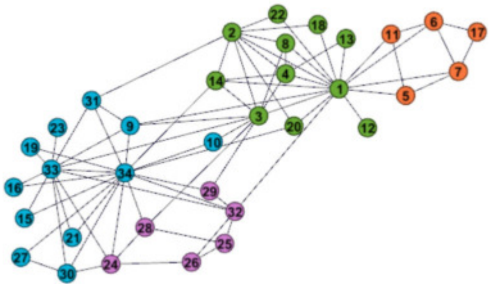

Fig. 1: The Zachary karate club network. There are 4 different communities denoted by different colors: *Green*, *Orange*, *Blue* and *Purple*.

Table 1: Causes for node $9 \notin Green$. ρ corresponds to responsibility, Q to modularity and γ to discrepancy. Not all nodes of *Green* are shown since they have exactly the same behavior.

Cause	Γ	ρ	Q	γ
$(9,12)$	$\{(9,2)\}$	$\frac{1}{2}$	0.402	0
$(9,20)$	$\{(9,2)\}$	$\frac{1}{2}$	0.402	0
$(9,14)$	$\{(9,2)\}$	$\frac{1}{2}$	0.402	0
$(9,4)$	$\{(9,2)\}$	$\frac{1}{2}$	0.402	0
$(9,8)$	$\{(9,2)\}$	$\frac{1}{2}$	0.402	0
$(9,2)$	$\{(9,8)\}$	$\frac{1}{2}$	0.402	0

contain the neighbours of 34, we could weaken node 34, by choosing some of its incident edges (with the exception of edge $(34,10)$), thus indirectly making node 10 belong to *Green*. However, this is not a cause for $10 \in Blue$ but a by-product of 34 being a hub node of *Blue*. Of course, by transitivity[4], the fact that 34 is a hub of *Blue* causes $10 \in Blue$ through edge $(34,10)$, but we prefer to look straightforwardly at the direct cause expressed by this edge.

Why does node $32 \notin Blue$? As seen in Figure 1, node 32 is quite central in *Purple* community. We found out that the edge $(31,32)$ is a cause for $32 \notin Blue$ with contingency $\Gamma = \{(19,32)\}$ meaning that its responsibility for $32 \notin Blue$ is $1/2$. Note that since the network is undirected the same can be said for edge $(19,32)$ as a cause with contingency $\Gamma = \{(31,32)\}$ with $\rho = 1/2$. In this case $Q = 0.404$ while node 29 is also put in *Blue* community, and thus $\gamma = 1/34$.

Finally, lets look at node 9. Why does $9 \notin Green$? Iterating over all nodes in *Green* as causes we get the results in Table 1. The results are expected since $N(9) = \{1,3,31,33,34\}$, which are the most central nodes w.r.t. degree in their communities. We expected that $(9,2)$ would be a counterfactual cause for $9 \notin Green$ but this is not the case.

What if we extend the definition of contingency and allow for deletions of edges of 9 to nodes within its current community so that its belongingness to *Blue* community is weakened? Then, in this case the edge $(9,34)$ is a cause with $\Gamma = \{(9,31)\}$ since their removal moves 9 to *Green* with $\gamma = 0$ and $Q = 0.423$.

[4] Transitivity does not hold in general w.r.t. causation [11].

3 Algorithmic Aspects

In this section we describe algorithmic aspects that allow us to answer why (Definition 1) and why-not (Definition 2) queries for community detection. We first provide a general framework that is oblivious to the community detection algorithm being used. Then, within this framework and for reasons of efficiency, we specialize by focusing on modularity-based algorithms.

3.1 The General Framework

We begin by describing a trivial algorithm-agnostic framework. In fact, this framework is so general that can be used as a first step for introducing causality in different network settings as argued in Section 4. Assume an algorithm $\mathcal{A}$ that divides a given network $G = (V, E)$ into a set of communities $\mathcal{C}$. We pose the question "why a node $v \in V$ belongs in community $C \in \mathcal{C}$" (henceforth *why question*). For the *why-not question* the framework works in the same way.

Following Definition 1, we need to identify edges within E_e that are causes and discover their respective contingency sets $\varGamma$ as well as the changes implied by them in the community structure in order to compute the responsibility ρ and the discrepancy γ. To accomplish this, we first iterate over all subsets c of E_e to choose possible causes e in increasing size (starting from singletons) and then we iterate on all subsets of E_e/e to compute $\varGamma$. We maintain the top-k causes with highest ρ. If we are interested on γ as well, we could use either a weighted mean or maintain the top-k dominating causes with respect to both metrics. A very crude upper bound for the method is $O(2^{2y})$ iterations of the algorithm $\mathcal{A}$, where $y = |E_e|$ is the number of endogenous pairs of nodes.

Apparently, the time complexity of this framework is prohibitive. To speed the algorithm up, we can use the Locality Assumption. In this sense, we can define the endogenous pairs of nodes to be all corresponding edges at a small distance from v. For example, if we include in the endogenous set the neighbours of v, then the number of iterations is $O(2^{2deg(v)})$, which is considerably smaller especially for sparse networks that are usually seen in practice. However, even in this case the number of iterations is quite large. We could further reduce the complexity by having some information about the inner workings of the algorithm $\mathcal{A}$. In the following, we assume such an approach by looking at an algorithm that optimizes modularity. In addition, for the *why question* we consider as endogenous pairs of nodes all the neighbours of the query node v.

3.2 Working with Modularity-based Methods

Firstly, we apply a modularity based community detection algorithm in the given network G, such as the Louvain method, which maximizes modularity. We refer the reader to [1] for more details. G is now partitioned into communities.

We focus on the *why question*. Subsequently we have to decide which edges will be examined as possible causes. Therefore, we use a combination of two metrics: *embeddedness* (ξ_v) and *degree* ($deg(v)$) of the query node v. The embeddedness ξ_v of v in community C, is defined as the ratio between the number of edges connecting v to nodes of C, and the degree of v [8], i.e., $\xi_v = deg^C(v)/deg(v)$. The higher the value of ξ_v, the stronger the belongingness of v to C.

However, this metric alone cannot be used in our case because it is misleading. Let's look at the example of Figure 1. The embeddedness of node 2 in the Green community is approximately equal to 0.89. On the other hand, the embeddedness of node 12 in the same community is equal to 1. However, node 12 has only one edge and it is rational for this edge to be incident to a node of the same community. Thus, we can combine embeddedness with the degree of the query node resulting in a metric M as follows: $M_v = \dfrac{\xi_v \cdot deg^C(v)}{\max(deg^C)}$, where $\max(deg^C)$ is the maximum node degree inside community C. As it can be understood, metric M is defined as above in order to reward edges that participate more actively in their community. It is also a simple metric that can be easily implemented. Note that instead of M, we can use any other metric. Consequently, we rank the edges by their M values in decreasing order, and consider as cause(s) the first x edge(s) of this ranking. The constant x is defined by the maximum number of causes as it has been assumed by the Bound Assumption. Then, we compute the corresponding ρ and γ values.

Now the structure of G has changed due to the interventions $\varDelta$, implied by the above causes and their $\varGamma$. Thus, we must apply again community detection in the new network G', which is G after the integration of $\varDelta$. As it may be inferred, the changes of G are not so radical and are observed to be around specific parts of G. For this reason, we can apply the Louvain method only to a part of G whose community affiliation might change due to the $\varDelta$. There are some incremental community detection approaches such as [13,30] that can be implemented along with either Louvain or any other modularity based community detection method.

4 Additional Issues and Extensions

In this section, we discuss various extensions to the framework discussed above.

Weighted Networks. The proposed approach can be extended to weighted networks as well. An undirected, weighted network $G = (V, E, w)$ is composed of a node set $V = \{1, ..., n\}$ with $n = |V|$ nodes and an edge set $E \subseteq V \times V$ with $m = |E|$ undirected edges and edge weights $w = E \rightarrow \mathbb{R}_{>0}$. Definitions 1 and 2 are straightforward to apply. The weighted responsibility is a simple extension of the unweighted case as we define $\varGamma = \sum_{e \in \varGamma} w(e)$.

In weighted networks, where the weight corresponds to how strongly two nodes are connected, the Locality Assumption implies that paths with large total additive weight are preferred over paths of lower weight. This is because it has been assumed that weights resemble similarity and not distance, in which case one has to consider the inverse of weights. The major difference is that in the unweighted case the choice for an edge is binary (remove/add). In the weighted case, the choice is not binary since the algorithms to identify causes must also be able to increase/decrease weights; e.g., in a social network these changes in weights may correspond to the strengthening/weakening of a friendship. This requires a strategy to handle these weights and affects the discrepancy measure.

Uncertain Networks. Our approach is naturally extended to the case of uncertain networks. An uncertain network $G = (V, E, P)$ is defined over a set of nodes V, a set of edges E between pairs of nodes and a probability distribution P over the edges E. Definitions 1 and 2 as well as ρ and γ can be straightforwardly generalized to uncertain networks, e.g., we can simply change the probability of existence of an edge and increase it or decrease it in order to prove actual causes. The approach will be very similar to the case of the weighted networks with additional restrictions related to handling probabilities.

Extending the Definition of Contingency. The contingency sets may be different considering the accepted actions we can do i.e. addition/removal of an edge, weight changes, etc. Note that in Definition 1, Γ contains edges to be removed from the network. Γ could also include, if necessary, edges to be added. Adding edges in this case strengthens the node's belongingness to other communities, thus moving it further away from community C. Similarly, in Definition 2, Γ contains edges to be added to the network. Γ could also include, if necessary, edges to be removed that could indirectly lead v to belong to another community. Besides, if the network is weighted, the contingency set may be altered if we consider the changes on the edges' weights. Although these extended definitions would provide more options, efficiency would be aggravated.

The Endogenous Pairs of Nodes. In general, for the *why question*, we can consider as endogenous any pair of nodes whose corresponding edges are incident to the query node. Furthermore, we can expand the former by adding edges that belong to the same community as the query node. For the *why-not question*, we can add to the endogenous set, pairs of nodes that belong to neighbouring communities of the query node's community. The choice of the endogenous pairs of nodes is of critical importance for the efficiency and depends heavily on the definitions of the actual cause and the contingency set as exemplified in our previous point. In addition, this choice affects how much freedom the algorithm will have in order to identify causes and especially surprising causes. A surprising cause would be a distant edge to node v whose removal would lead v to change its community according to algorithm $\mathcal{A}$. This trade-off needs to be handled carefully, as the more wider the endogenous sets, the more the availability of possible causes but at the same time the more processing time will be needed.

Beyond Community Detection. The general but inefficient framework proposed in Section 3, can be readily extended to other network-related processing problems as well. For example, asking why a node belongs to the k-core of the network [17] can be tackled by this framework given that the appropriate changes have been made to Definitions 1-4. In the following, we show how Definition 1 would change in this case.

Definition 5. *Let $e \in E_e$ be an endogenous pair of nodes and let v belong to the k-core.*

- *e is called a counterfactual cause for v in the k-core if for the network $G = (V, E)$ it holds that v belongs to the k-core while for $G' = (V, E - \{e\})$ it holds that v does not belong to the k-core.*

– *e is called an actual cause for v belonging in the k-core if there exists a set of edges $\Gamma \subseteq E_e$ called a contingency for e such that e is a counterfactual cause for v belonging in the k-core in the network $G' = (V, E - \Gamma)$.*

In this case, E_e could contain all edges in the k-core of the network since these are the possible causes for node v being in the k-core. Similarly, one can introduce causality in the minimum cut problem in a weighted network (why does edge e belong to the cut?). Efficiency issues must be handled in an ad-hoc manner based on the problem at hand.

In the present work, we have introduced the concept of causal explanations in community formation and we have proposed a framework for identifying actual causes. In the future, we will focus on efficient algorithmic techniques as well as on extensive experimental evaluation for different types of networks (e.g., directed, weighted) and different problems (e.g., overlapping communities, k-core decomposition).

Ευρωπαϊκή Ένωση
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgements. Georgia Baltsou is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research - 2nd Cycle" (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (oct 2008)
2. Carmi, E., Oestreicher-Singer, G., Sundararajan, A.: Is oprah contagious? identifying demand spillovers in online networks. Identifying Demand Spillovers in Online Networks (Aug, 2012) .NET Inst. Working Paper (10-18) (2012)
3. Chen, S., Wang, Z.Z., Tang, L., Tang, Y.N., Gao, Y.Y., Li, H.J., Xiang, J., Zhang, Y.: Global vs local modularity for network community detection. PloS one 13(10), e0205284 (2018)
4. Chockler, H., Halpern, J.Y.: Responsibility and blame: A structural-model approach. J. of Artificial Intelligence Research 22(1), 93–115 (2004)
5. De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: Mixing local and global information for community detection in large networks. J. of Computer and System Sciences 80(1), 72–87 (2014)
6. De Meo, P., Ferrara, E., Fiumara, G., Ricciardello, A.: A novel measure of edge centrality in social networks. Knowledge-based systems 30, 136–150 (2012)
7. Derrible, S., Kennedy, C.: Network analysis of world subway systems using updated graph theory. Transportation Research Record 2112(1), 17–25 (2009)
8. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics reports 659, 1–44 (2016)
9. Friedkin, N.E.: Horizons of Observability and Limits of Informal Control in Organizations*. Social Forces 62(1), 54–77 (09 1983)
10. Gao, Y., Liu, Q., Chen, G., Zhou, L., Zheng, B.: Finding causality and responsibility for probabilistic reverse skyline query non-answers. IEEE Transactions on Knowledge and Data Engineering 28(11), 2974–2987 (Nov 2016)

11. Halpern, J.Y., Pearl, J.: Causes and Explanations: A Structural-Model Approach. Part I: Causes. The British J. for the Phil. of Science 56(4), 843–887 (12 2005)
12. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. part ii: Explanations. The British J. for the Phil. of Science 56(4), 889–911 (2005)
13. Held, P., Krause, B., Kruse, R.: Dynamic clustering in social networks using louvain and infomap method. In: ENIC. pp. 61–68 (2016)
14. Javed, M.A., Younis, M.S., Latif, S., Qadir, J., Baig, A.: Community detection in networks: A multidisciplinary review. J. of Network and Computer Applications 108, 87–111 (2018)
15. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics 6(1), 29–123 (2009)
16. Lian, X., Chen, L.: Causality and responsibility: Probabilistic queries revisited in uncertain databases. p. 349–358. CIKM '13, New York, NY, USA (2013)
17. Malliaros, F.D., Giatsidis, C., Papadopoulos, A.N., Vazirgiannis, M.: The core decomposition of networks: theory, algorithms and applications. VLDB Journal 29(1), 61–92 (2020)
18. Meliou, A., Gatterbauer, W., Moore, K.F., Suciu, D.: The complexity of causality and responsibility for query answers and non-answers. Proc. VLDB Endow. 4(1), 34–45 (Oct 2010)
19. Meliou, A., Roy, S., Suciu, D.: Causality and explanations in databases. Proceedings of the VLDB Endowment 7, 1715–1716 (08 2014)
20. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. pp. 29–42 (2007)
21. Newman, M.E.: Modularity and community structure in networks. Proceedings of the national academy of sciences 103(23), 8577–8582 (2006)
22. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
23. Rossetti, G., Cazabet, R.: Community discovery in dynamic networks: A survey. ACM Comput. Surv. 51(2) (Feb 2018)
24. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular systems biology 3, 88 (02 2007)
25. Shen, H.W.: Community structure of complex networks. Springer Science & Business Media (2013)
26. Tripathi, B., Parthasarathy, S., Sinha, H., Raman, K., Ravindran, B.: Adapting community detection algorithms for disease module identification in heterogeneous biological networks. Frontiers in genetics 10, 164 (2019)
27. Wang, Z., Wang, C., Ye, X., Pei, J., Li, B.: Propagation history ranking in social networks: A causality-based approach. Tsinghua Science and Tech. 25(2), 161–179 (2020)
28. Wilber, A.W., Doye, J.P., Louis, A.A., Lewis, A.C.: Monodisperse self-assembly in a model with protein-like interactions. J. of chemical physics 131(17), 11B602 (2009)
29. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. of anthropological research 33(4), 452–473 (1977)
30. Zarayeneh, N., Kalyanaraman, A.: A fast and efficient incremental approach toward dynamic community detection. In: Proc. of the ACM/IEEE Int. Conference on Advances in Social Networks Analysis and Mining. pp. 9–16 (2019)