Rozenshtein, Polina; Tatti, Nikolaj; Gionis, Aristides

Inferring the strength of social ties

Published in:
KDD 2017 - Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

DOI:
10.1145/3097983.3098199

Published: 13/08/2017

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Rozenshtein, P., Tatti, N., & Gionis, A. (2017). Inferring the strength of social ties: A community-driven approach. In *KDD 2017 - Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. Part F129685, pp. 1017-1025). ACM. https://doi.org/10.1145/3097983.3098199

Inferring the Strength of Social Ties: A Community-Driven Approach

Polina Rozenshtein
HIIT, Aalto University
Espoo, Finland
polina.rozenshtein@aalto.fi

Nikolaj Tatti
HIIT, Aalto University
Espoo, Finland
nikolaj.tatti@aalto.fi

Aristides Gionis
HIIT, Aalto University
Espoo, Finland
aristides.gionis@aalto.fi

ABSTRACT

Online social networks are growing and becoming denser. The social connections of a given person may have very high variability: from close friends and relatives to acquaintances to people who hardly know. Inferring the strength of social ties is an important ingredient for modeling the interaction of users in a network and understanding their behavior. Furthermore, the problem has applications in computational social science, viral marketing, and people recommendation.

In this paper we study the problem of inferring the strength of social ties in a given network. Our work is motivated by a recent approach [27], which leverages the *strong triadic closure* (STC) principle, a hypothesis rooted in social psychology [13]. To guide our inference process, in addition to the network structure, we also consider as input a collection of *tight* communities. Those are sets of vertices that we expect to be connected via strong ties. Such communities appear in different situations, e.g., when being part of a community implies a strong connection to one of the existing members.

We consider two related problem formalizations that reflect the assumptions of our setting: small number of STC violations and strong-tie connectivity in the input communities. We show that both problem formulations are **NP**-hard. We also show that one problem formulation is hard to approximate, while for the second we develop an algorithm with approximation guarantee. We validate the proposed method on real-world datasets by comparing with baselines that optimize STC violations and community connectivity separately.

CCS CONCEPTS

•Information systems → Collaborative and social computing systems and tools; Data mining; •Theory of computation → Graph algorithms analysis;

KEYWORDS

Social network analysis, strong triadic closure, network inference, approximation algorithms

KDD'17, August 13–17, 2017, Halifax, NS, Canada.
 978-1-4503-4887-4/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3097983.3098199

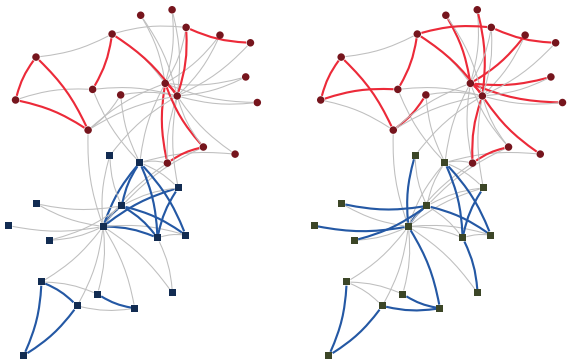

Figure 1: Strong edges in the Karate-club dataset inferred by the algorithm of Sintos and Tsaparas [27] (left) and our method (right) using two teams. The colors of the edges and the vertices depict the two teams.

1 INTRODUCTION

The growth of online social networks has been an important factor in shaping our lives for the 21st century. 68 % of adults in the US, also accounting for those who do not use Internet at all, are Facebook users.[1] Over the past few years, an ecosystem of online social-network platforms has emerged, serving different needs and purposes: being connected with close friends, sharing news and being informed, sharing photos and videos, making professional connections, and so on.

The emergence of such social-networking platforms has introduced many novel research directions. First, online systems have enabled recording and studying human behavior at a very large scale. Second, the specific features of the different systems are changing the way people interact with each other: new social norms are formed and human behavior is adapting. Consequently, data collected by online social-network systems are used to analyze and understand human behavior and complex social phenomena. Questions of interest include understanding information-diffusion phenomena, modeling network evolution and predicting future behavior, identifying the role of users and network links, and more.

A question of particular importance, which is the focus of this paper, is the problem of inferring *the strength of social ties* in a network. Quantifying the strength of social ties is an essential task for sociologists interested in understanding complex network dynamics based on pair-wise interactions [13], or for engineers interested in designing applications related to viral marketing [7] or friend recommendation [21].

[1] http://www.pewinternet.org/2016/11/11/social-media-update-2016/

The problem of inferring the strength of social ties in a network has been studied extensively in the graph-mining community [11, 12, 25, 27–29]. While most approaches use user-level features in order to estimate the social-tie strength between pairs of users, our approach, inspired by the work of Sintos and Tsaparas [27], relies on the *strong triadic closure* (STC) principle [5, 8, 13]. The STC principle assumes that there are two types of ties in the social network: *strong* and *weak*. It then asserts that it is unlikely to encounter a triple of users so that two of the ties are strong while the third is missing. In other words, two users who have a strong tie to a third common friend should be acquainted to each other, i.e., they should have *at least* a weak tie to each other.

Sintos and Tsaparas [27] address the problem of inferring the strength of social ties (i.e., labeling the links of a given network as *strong* or *weak*) by leveraging the STC property in an elegant manner. They first assume that users are more interested in establishing and maintaining strong ties, as presumably, this is the reason that they joined the network. Using this assumption they formulate the link-strength inference problem by asking to assign the maximal number of strong ties (or the minimal number of weak ties) so that the STC property holds. They prove that the problem is **NP**-hard and they devise an approximation algorithm for the variant of minimizing the number of weak ties.

In this paper, in addition to the network structure, we also consider as a collection of topical communities $C_1, \ldots, C_k$. We assume that the given communities are *tight*, that is, each community C_i represents a set of users with *focused* interest at a particular topic. For example, such a tight community may be (*i*) a set of users who have been actively involved in a discussion in the social network about a certain issue, (*ii*) the set of scientists who work on 'deep learning,' or (*iii*) the HR team of a company.

We then require that each given community C_i should be connected via strong ties. In other words, for every two nodes in C_i there is a path made of *strong* ties. This requirement reflects the fact that we consider tight communities, as the examples above. Clearly this constraint is less meaningful if we consider *loose* communities, i.e., all facebook users who like the 'Friends' TV series.

Equipped with these assumptions we now define the problem of inferring the strength of social ties: given a social network $G = (V, E)$, and a set of tight communities $C_1, \ldots, C_k \subseteq V$, we ask to label all the edges in E as either *strong* or *weak* so that (*i*) each community C_i is connected via strong ties; and (*ii*) the total number of STC violations is minimized. Our problem definition captures two natural phenomena: first, tight communities tend to have a backbone, e.g., being part of a community implies a strong connection to one of the existing members. Second, strong ties tend to close triangles, as postulated by the strong triadic closure principle, and thus, real-world social networks have relatively few STC violations. We call a triple of nodes, interconnected by 3 edges, a *closed* triangle. A triple with only 2 edges in common is called an *open* triangle.

Example. An illustration of our method on the Karate-club dataset [30] is shown in Figure 1. Our method (right) is contrasted with the algorithm of Sintos and Tsaparas [27] (left). Both approaches use the STC principle, but additionally, our method requires that certain communities provided as input are connected with strong ties. In the example, we consider the two ground-truth communities of the Karate-club dataset. We observe that the sets of strong ties inferred by the two methods are fairly similar. We also observe that our method introduces an STC violation only when it is necessary for ensuring connectivity. On the other hand, the method of Sintos and Tsaparas [27] leaves several disconnected singleton nodes, which is less intuitive. □

We capture the above intuition using two related problem definitions. For the first problem (MinViol) we ask to minimize the number of STC violations, while for the second problem (MaxTri) we ask to maximize the number of non-violated open triangles — there cannot be a violation on a closed triangle. In both cases we label the network edges so as to satisfy the connectivity constraint, with respect to strong edges, for all input communities.

We show that both problems, MinViol and MaxTri, are **NP**-hard, even if the input consists of one community. Furthermore, we show that MinViol is hard to approximate to any multiplicative factor. On the other hand, the problem MaxTri is amenable to approximation: its objective function is submodular and non-decreasing, while the connectivity constraints can be viewed as an intersection of matroids. Thus, the classic result of Fisher et al. [10] applies, implying that a greedy algorithm leads to $1/(k+1)$ approximation ratio.

We evaluate our methods on real-world networks and input communities. Our quantitative results show that our method achieves a balance between baselines that optimize STC violations and community connectivity separately, while our case study suggests the strong edges selected by the method are meaningful and intuitive.

The remaining paper is as follows. We introduce the notation and give the problem definition in Section 2. We show the computational hardness in Section 3 and present the approximation algorithm in Section 4. The related work is discussed in Section 5, and the experimental evaluation is given in Section 6. We conclude the paper with remarks in Section 7.

2 PRELIMINARIES AND PROBLEM DEFINITION

The main input for our problem is an undirected graph $G = (V, E)$ with n vertices and m edges. Given a subset of vertices $X \subseteq V$ and a subset of edges $F \subseteq E$, we write $F(X)$ to denote the edges in F that are connected to vertices in X.

We are interested in labeling the set of edges E. Specifically, we want to label each edge as either *strong* or *weak*.

To specify a labeling of edges E it is sufficient to specify the set of strong edges $S \subseteq E$. To quantify the quality of a labeling $S \subseteq E$, for a given graph $G = (V, E)$, we use the *strong triadic closure* (STC) property. Namely, given a triple (u, v, w) of vertices such that $(u, v), (v, w) \in S$, we say that the triple violates the STC property if $(u, w) \notin E$. In other words, a

strong friend of a strong friend must be connected, possibly with a weak edge. We define $viol(S; G)$ to be the number of STC violations. Typically, G is known from the context, and we omit it from the notation.

As described in the introduction, our goal is to discover a strong backbone of the graph. At simplest we are looking for a set of edges that connect the whole graph with strong ties while minimizing the number of violations.

We also consider the more general setup, where are given a *set of communities* (possibly overlapping), each community is simply a subset of vertices, and the goal is to make sure that each community is connected on its own with strong ties, without using outside edges.

More formally, we have the following problem definition

Problem 1 (MinViol). *Given a graph $G = (V, E)$ and a set of communities $C_1, \ldots, C_k \subseteq V$, find a set of strong edges $S \subseteq E$ such that each $(C_i, S(C_i))$ is connected and the number of* STC *violations, $viol(S)$, is minimized.*

In the above problem definition $(C_i, S(C_i))$ is the subgraph of G induced by the vertices in C_i *and* the edges in S, that is, $S(C_i) = \{(u, v) \in E \mid u, v \in C_i \text{ and } (u, v) \in S\}$.

In order for MinViol to have at least one feasible solution, we assume that $(C_i, E(C_i))$ is connected for each C_i.

In addition to minimization version, we consider a maximization version of the problem. In order to do that, given a graph G, let T be the number of open triangles in G. We define $tri(S) = T - viol(S)$ to be the number of open triangles that are not violated.

This leads to the following optimization problem.

Problem 2 (MaxTri). *Given a graph $G = (V, E)$ and a set of communities $C_1, \ldots, C_k \subseteq V$, find a set of strong edges S such that each $(C_i, S(C_i))$ is connected, and the number of non-violated triangles, $tri(S)$, is maximized.*

Note that STC violations can occur only for open triangles. Therefore, $tri(S) = T - viol(S)$ is nonnegative, while it achieves its maximum value T when there are no STC violations.

Obviously, MinViol and MaxTri have the same optimal answer. However, we will see that they yield different approximation results: MinViol cannot have any multiplicative approximation guarantee (constant or non-constant) while a greedy algorithm has $1/(k+1)$ guarantee for MaxTri.

3 COMPUTATIONAL COMPLEXITY

Our next step is to establish that MinViol (and MaxTri) are **NP**-hard. Moreover, we show that MinViol cannot have any multiplicative approximation guarantee.

Proposition 3.1. *Deciding whether there is a solution MinViol with zero violations is* **NP***-complete. Thus, there is no multiplicative approximation algorithm for MinViol, unless* **P**=**NP**. *The result holds even if we use only one community.*

Proof. To prove the result we will reduce CliqueCover to MinViol. In an instance of CliqueCover, we are asked to partition a graph $G = (V, E)$ to k subgraphs, each one of them being a clique.

Assume a graph $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$, and an integer k.

For the reduction of CliqueCover to MinViol, we first define a graph $H = (W, A)$. The vertex set W consists of $2n + k$ vertices grouped in 3 sets: the first set are the original vertices V, the second set is U with n vertices, the third set is X containing k vertices.

The edges A are as follows: We keep the original edges E. For each $i = 1, \ldots, n$ and $j = 1, \ldots, k$, we add (u_i, v_i), (v_i, x_j), and (u_i, x_j). Here $v_i \in V$, $u_i \in U$ and $x_j \in X$. We also fully-connect X.

We add one community consisting of the whole graph.

We claim that there is a 0-solution to MinViol if and only if there is a clique cover for G. Since CliqueCover is **NP**-hard, this automatically proves the inapproximability.

Assume first that we are given a clique cover $\mathcal{P} = \{P_1, \ldots, P_k\}$. Define the following set of strong edges. For each vertex v_i, let P_j be the clique containing v_i; add edges (u_i, v_i), (v_i, x_j) to S. Finally, add an edge (x_1, x_j) for each $j = 2, \ldots, k$. It is straightforward to see that the connectivity constraints are satisfied. The strong wedges are

$$
\begin{aligned}
&u_i\text{–}v_i\text{–}x_j, && \text{for} \quad v_i \in P_j,\\
&v_i\text{–}x_j\text{–}x_1, && \text{for} \quad v_i \in P_j, \text{ and } j \neq 1,\\
&v_i\text{–}x_1\text{–}x_j, && \text{for} \quad v_i \in P_1, \text{ and } j \neq 1,\\
&x_j\text{–}x_1\text{–}x_q, && \text{for} \quad q \neq j,\\
&v_i\text{–}x_j\text{–}v_\ell, && \text{for} \quad v_i, v_\ell \in P_j, \text{ and } i \neq \ell.
\end{aligned}
$$

None of these wedges induce a violation, the last one follows from the fact that $\mathcal{P}$ is a clique cover. Thus $viol(S) = 0$.

To prove the other direction, let S be the set of strong edges such that $viol(S) = 0$.

Let $i = 1, \ldots, n$. To satisfy the connectivity, $(u_i, x_j) \in S$ or $(u_i, v_i) \in S$ (or both) for some j. Define $Y = \left\{v_i; (u_i, x_j) \in S\right\}$ and $Z = V \setminus Y$.

Let $v_i \in Z$. Since $(u_i, v_i) \in S$, all edges adjacent to v_i in E are weak. Thus, to satisfy the connectivity, we must have $(v_i, x_j) \in S$.

Define two families $\mathcal{A}$ and $\mathcal{B}$, each of k sets, by

$$
\begin{aligned}
A_j &= \left\{v_i \in Y; (u_i, x_j) \in S\right\} \quad \text{and}\\
B_j &= \left\{v_i \in Z; (v_i, x_j) \in S\right\}.
\end{aligned}
$$

Write $A_0 = B_0 = \emptyset$, and define a family $\mathcal{P}$ of k disjoint sets by $P_j = (A_j \cup B_j) \setminus P_{j-1}$. $\mathcal{P}$ covers V since each vertex in V is in A_j or B_j for some j.

We claim that $\mathcal{P}$ is a clique cover. To see this, let $v_i, v_\ell \in P_j$. If $v_i \in Y$ and $v_\ell \in Z$, then $u_i\text{–}x_j\text{–}v_\ell$ is a violation since $i \neq \ell$. If $v_i, v_\ell \in Y$, then $u_i\text{–}x_j\text{–}u_\ell$ is a violation, or $i = \ell$. If $v_i, v_\ell \in Z$, then either $i = \ell$ or $(v_i, v_\ell) \in E$. This shows that $\mathcal{P}$ is a clique cover. □

Corollary 3.2. *The MaxTri problem is* **NP***-hard. The result holds even if we use only one community.*

Algorithm 1: Greedy algorithm for MaxTri

```
S ← E; A ← E;
while A ≠ ∅ do
    e = arg max_{e∈A} tri(S \ {e});
    if S \ {e} satisfies the connectivity constraints then
        S ← S \ {e};
    A ← A \ {e};
return S;
```

4 APPROXIMATION ALGORITHM

In the previous section we saw that the problems MinViol and MaxTri are **NP**-hard, even for one community, and additionally, MinViol is hard to approximate to any multiplicative factor. In this section we show that MaxTri can be approximated with $1/(k+1)$ guarantee, where k is the number of communities in the input. As an important consequence, if we have one community, we can find a solution with approximation guarantee $1/2$. Furthermore, it follows that if all communities are edge-disjoint, our algorithm yields a $1/2$ approximation guarantee.

To prove the approximation algorithm we argue that $tri(\cdot)$ is submodular with respect to weak edges. Moreover, the connectivity constraint of each community can be viewed as a matroid. Thus, satisfying all the connectivity constraints is an intersection of matroids.

These properties allow us to use a classic result of maximizing a submodular function over an intersection of k matroids: Fisher et al. [10] showed that a greedy algorithm leads to $1/(k+1)$ approximation ratio. Here the greedy algorithms starts with none of the edges being weak, that is, all edges are strong. We find a strong edge, say e, inducing the most violations. We convert e to a weak edge if the connectivity constraints allow it. Otherwise, we let e being strong. The pseudo-code is given in Algorithm 1.

Note that our problem formulation is agnostic with respect to whether strong edges should be maximized or minimized. In Algorithm 1 strong edges are kept, even if they are not crucial for connectivity, as long as they do not induce any violations. This behavior is in line with the idea of Sintos and Tsaparas [27], who aim to maximize the number of strong edges. It is in contrast, however, with our second baseline, the algorithm of Angluin et al. [1], who want to find a minimum set of edges to ensure connectivity. If we wish to obtain a minimal number of strong edges, we can continue the main iteration in Algorithm 1 and convert to weak all edges that are not necessary for connectivity and do not create any STC violations.

We now show the properties required by the result of Fisher et al. [10] for the greedy algorithm to yield approximation ratio $1/(k+1)$. We first show that the function $tri(\cdot)$ is submodular with respect to weak edges.

Proposition 4.1. *Consider a graph $G = (V, E)$. Let $f(W) = tri(E \setminus W)$. The function f is submodular and non-decreasing.*

Proof. To prove the submodularity we show that $viol(\cdot)$ is supermodular with respect to strong edges. This makes $tri(\cdot)$ submodular with respect to strong edges, which in turn makes f submodular with respect to weak edges. For the last implication it is well-known that a function is submodular if and only if its complement is submodular.[2]

Let S be a set of strong edges. For a vertex u, define $N_X(u) = \{v \mid (u,v) \in S\}$ to be the strong-neighbors of u. Define also $\overline{N}(u) = \{v \mid (u,v) \notin E\}$ to be the non-neighbors or u.

The number of additional STC violations introduced by labeling edge $e = (u,v)$ as strong is

$$viol(S \cup \{e\}) - viol(S) = \left|N_S(u) \cap \overline{N}(v)\right| + \left|N_S(v) \cap \overline{N}(u)\right|.$$

Let $T \subseteq S$. For any $u \in V$ and any edge set W, we have

$$|N_T(u) \cap W| \leq |N_S(u) \cap W|. \tag{1}$$

This implies that for $T \subseteq S \subseteq E$ and any edge $e \notin S$,

$$viol(T \cup \{e\}) - viol(T) \leq viol(S \cup \{e\}) - viol(S),$$

which proves the supermodularity of $viol(\cdot)$.

To prove the monotonicity, we show that $viol(\cdot)$ is non-decreasing. This makes $tri(\cdot)$ non-increasing, which makes f non-decreasing.

Consider any subset $Q \subseteq S$ of strong edges. Consider any triangle that violates STC. It must be that two of the edges are strong and the third one is missing. When we add more strong ties to S, this violating triangle would still violating STC, so the number of violating triangles does not decrease. Consequently, $viol(\cdot)$ is non-decreasing. □

Our next step is to argue that the connectivity constraints are matroids with respect to weak edges. Fortunately, this is a known result and these matroids are commonly known as *bond matroids*, see for example, Proposition 3.3 by Oxley [26].

Proposition 4.2. *Assume a graph $G = (V, E)$ and a subset C such that $(C, E(C))$ is connected. Define a family of sets*

$$\mathcal{M} = \{W \subseteq E; (C, E(C) \setminus W(C)) \text{ is connected}\}.$$

Then $\mathcal{M}$ is matroid.

The two propositions show that we can use the result by Fisher et al. [10], and obtain $1/(k+1)$ guarantee, where k is the number of communities, i.e., the sets of vertices for which we require a connectivity constraint.

We can obtain a better guarantee, $1/2$, if we know that communities are edge-disjoint. This follows from the fact that we can express the connectivity constraints as a single matroid.

Proposition 4.3. *Assume a graph $G = (V, E)$ and family of subsets $C_1, \ldots, C_k$, such that induced subgraphs $(C_i, E(C_i))$ are edge-disjoint and each $(C_i, E(C_i))$ is connected. Define a family of sets*

$$\mathcal{M} = \{W \subseteq E; (C_i, E(C_i) \setminus W(C_i)) \text{ is connected, for every } i\}.$$

Then $\mathcal{M}$ is matroid.

The result follows from immediately Proposition 4.2. and the following standard lemma which we state without a proof.

[2] see, for example, http://melodi.ee.washington.edu/~bilmes/ee595a_spring_2011/lecture1_presented.pdf for a proof.

Lemma 4.4. *Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be k matroids, each matroid $\mathcal{M}_i$ is defined over its own ground set U_i. Then a direct sum*

$$\mathcal{M} = \left\{ \bigcup_{i=1}^{k} X_i \mid X_i \in \mathcal{M}_i \right\}$$

is a matroid over $\bigcup_{i=1}^{k} U_i$.

Thus we can use the result by Fisher et al. [10] but now we have only one matroid instead of k matroids. This gives us an approximation guarantee of $1/2$.

Computational complexity: Let us finish with the computational-complexity analysis of the greedy algorithm. Assume that given a graph $G = (V, E)$, we have already enumerated all open triangles. Let t be the number of such triangles.

During the while-loop of the greedy algorithm, we maintain a priority queue for m edges, prioritized with the number of violations induced by a single edge. Whenever, a strong edge is deleted, we visit every open triangle induced by this edge and reduce the number of violations of the strong sister edge by 1. Note that we visit every triangle at most twice, so maintaining the queue requires $O(t + m \log n)$ time, if we use Fibonacci heap. To check the connectivity, we can use the technique introduced by Holm et al. [14], allowing us to do a connectivity check in $O(\log^2 n)$ amortized time. Thus, in total we need $O(t + km \log^2 n)$ time, plus the time to build the list of open triangles. Building such a list can be done in $O\left(\sum_v \deg^2(v)\right)$ time.

5 RELATED WORK

The study of interpresonal ties has a long history in social psychology. Several researchers have investigated the role of different types of social ties with respect to structural properties of social networks, as well as with respect to information-propagation phenomena. For example, in economics, Montgomery [23] showed that weak ties are positively correlated to higher wages and higher aggregate employment rates. More recent works considered how different social ties are formed and how they evolve in online social networks, such as email networks [17] and mobile-phone networks [25].

The strong triadic closure (STC) property, which forms the basis of our inference algorithm, was first formulated in the seminal paper of Granovetter [13], while evidence that this property holds in social networks has appeared in earlier works [5, 24]. Memic [22] have conducted a more recent study confirming that the principle remains valid on more recently-collected datasets [22].

In computer science, there have been works that study the problem of inferring the strength of social ties in a network.

Kahanda and Neville [16] use transactional events, such as communication and file transfers, to predict link strength, by applying techniques from the literature of the link-prediction problem [20]. It is shown that the approach can accurately predict strong relationships. Gilbert and Karahalios [12] propose a predictive model for inferring tie strength. The model uses variables describing the interaction of users in a social-media platform. The paper also illustrates how the inferred tie strength can be used to improve social-media features, such as, privacy controls, message routing, and friend recommendation. Likewise, Xiang et al. [29] leverage user-level interaction data. They formulate the problem of inferring hidden relationship strengths using a latent-variable model, which is learned by a coordinate-ascent optimization procedure. A feature of their setting is that social strengths are modeled as real-valued variables, not just binary. Jones et al. [15] examined a large set of features for the task of predicting the strength of social ties on a Facebook interaction dataset, and found that the frequency of online interaction is diagnostic of strong ties, while private communications (messages) are not necessarily more informative than public communications (comments, wall posts, and other interactions).

Backstrom and Kleinberg [2] consider a particular type of social ties — romantic relationships — and they ask whether this can be accurately recognized. They use a large sample of Facebook data to answer the question affirmatively, and on the way they develop a new type of tie strength, *the extent to which two people's mutual friends are not themselves well-connected*, which they call "dispersion."

In a different direction, Fang and Tang [9] consider only closed triangles and ask whether it is possible to find out which edges are formed last, i.e., which edges closed an open triad. The underlying research question is to recover the dynamic information in the triadic-closure process. They approach this problem using a probabilistic factor-graph model, and apply the proposed model on a large collaboration network.

Researchers have also studied the tie-strength inference problem in the presence of more than one social network. Gilbert [11] explore how well a tie strength model developed for one social-media platform adapts to another, while Tang et al. [28] consider a generalization of the problem over multiple heterogeneous networks. Their work uses a transfer-based factor-graph model, and also incorporates features motivated from social-psychology theories, such as social balance [8], structural holes [3], and and social status [6].

Most of the above works on tie-strength inference utilize pairwise user-level interaction data, such as email, private messages, public mentions, frequency of interactions, etc. In many cases such detailed data are not available. Our objective is to address the tie-strength inference problem using non-private data, such as the structure of the social network and information about communities that users have participated.

Conceptually and methodologically our paper is related to the work of Sintos and Tsaparas [27], who use as available information only the network structure, to infer strong and weak ties with the means of the strong triadic closure property [8]. We extend that work by introducing community-level information and a corresponding connectivity constraint to account for explaining the observed community structure: namely, we require that each community should be connected via strong ties. Like the work of Sintos and Tsaparas [27], we follow a combinatorial approach, but the techniques we use are significantly different.

A problem related to the inference of tie strength is the problem of predicting edge signs in social networks [4, 18]. The sign of an edge is typically interpreted as 'friend' or 'foe',

and thus, existing algorithms utilize theories from social psychology that are developed for this kind of relationships, in particular social balance [8] and social status theory [6].

In our experiments we are comparing our method with the algorithm of Angluin et al. [1], which takes as input a set of teams (communities) over a set of entities and seeks to add a minimal number of edges among the entities so that all given teams are connected. The algorithm is greedy and it is shown to have a $O(\log n)$ approximation guarantee. In our case, in addition to the set of teams we also have as input an underlying network, and edges are selected only if they are network edges. Selected edges are considered strong, and non-selected edges are considered weak. Thus, the method of Angluin et al. [1] is a combinatorial approach, which aim to satisfy connectivity among the input communities (like our method), but it does not take into account STC violations. On the other hand, it aims to minimize the number of strong edges (while our method is oblivious to this consideration).

6 EXPERIMENTAL EVALUATION

In this section we present our experimental evaluation. We describe the used datasets, discuss the baselines, and then present results of quantitative experiments and a case study.

The datasets and the implementation of the methods used in our experimental evaluation are publicly available.[3]

Datasets. We use 10 datasets, each dataset consists of a network and a set of communities. We describe these datasets below, while their basic characteristics are shown in Table 1. To ensure connectivity of each community, we selected only one part of each disconnected community, which induces the largest connected component.

• *KDD* and *ICDM* are subgraphs of the DBLP co-authorship network, restricted to articles published in the respective conferences. Edges represent co-authorships between authors. Communities are formed by keywords from paper abstracts.

• *FB-circles* and *FB-features* are Facebook ego-networks available at the SNAP repository [19]. In *FB-circles* the communities are social-circles of users. In *FB-features* communities are formed by user profile features.

• *lastFM-artists* and *lastFM-tags* are friendship networks of last.fm users.[4] A community in *lastFM-artists* and *lastFM-tags* is formed by users who listen to the same artist and genre, respectively.

• *DB-bookmarks* and *DB-tags* are friendship networks of Delicious users.[5] Community in *DB-bookmarks* and *DB-tags* is formed by users who use the same bookmark and keyword, respectively.

Additionally, we use SNAP datasets [19] with ground-truth communities. To have more focused groups, we only keep communities with size less than 10. To avoid having disjoint communities, we start from a small number of seed communities and iteratively add other communities that intersect at least one of the already selected. We stop when the number of vertices reaches 10 000. In this way we construct the following datasets:

• *DBLP*: This is also a co-authorship network. Communities are defined by publication venues.

• *Youtube*: This is a social network of Youtube users. Communities consist of user groups created by users.

Baselines. There are no direct baselines to our approach since the problem definition is novel. Instead we focus on comparing our method with the two techniques that inspired our approach. The first method is by Sintos and Tsaparas [27] and it maximizes the number of strong edges while keeping the number of STC violations equal to 0. The second method is by Angluin et al. [1] and it minimizes the number of edges needed to connect the communities. We refer to these methods as *Sintos* and *Angluin*, respectively, and we call our method *Greedy*.

Note that *Angluin* is oblivious to the STC property while *Sintos* does not use any community information. As we combine both goals, we expect that *Greedy* results in a compromise of these two baselines.

Comparison with the baselines. We compare the performance of *Greedy* with the two baselines. We run all algorithms on our datasets and measure the number of edges selected as strong, the number of STC violations, and the number of connected components created by strong edges for each of the input communities (so as to test the fragmentation of the communities). The results are shown in Table 2. The number of STC violations and the number of strong edges are reported as ratios (see table) for easy comparison. As expected, *Angluin* introduces more STC violations than *Greedy*: typically between 1%–21%. Interestingly, *Angluin* introduces less violations in *lastFM-tags* and 60 times more violations in *lastFM-artists*.

On the other hand, *Sintos* results in disconnected communities, ranging from 1.74 to 8.76 connected components per community, on average.

In the second experiment we test whether strong and weak ties can predict intra- and inter-community edges, respectively. The rationale of this experiment is to test the hypothesis that weak ties are bridges between different communities. With respect to the different methods, our objective is to further demonstrate that *Greedy* method results as a middle ground between *Angluin* and *Sintos*.

We randomly select half of the communities as *test communities* and run *Greedy* and *Angluin* using as input the underlying network and the other half of the communities. We also run *Sintos* using as input the underlying network; recall that this algorithm does not use any community information as input. Next, using the test communities we construct a set of intra-community edges E_{intra}, consisting of edges that belong to at least one community, and inter-community edges E_{inter}, consisting of edges that bridge two communities (but do not belong to any single community).

Let us denote all strong edges in the output of a given method as S and the weak edges as W. We define precision P_W and recall R_W for weak edges as

$$P_W = \frac{|W \cap E_{inter}|}{|W|} \quad \text{and} \quad R_W = \frac{|W \cap E_{inter}|}{|E_{inter}|},$$

[3] https://github.com/polinapolina/connected-strong-triadic-closure
[4] grouplens.org/datasets/hetrec-2011/
[5] www.delicious.com

Table 1: Network characteristics. $|V|$: number of vertices; $|E|$: number of edges in the underlying network; $|V_0|$: number of vertices, which participate in any given set (community); $|E_0|$: number of edges induced by communities; ℓ: number of sets (communities); $\text{avg}(\alpha_0)$: average density of subgraphs induced by input communities; $s_{\min}$, s_{avg}: minimum and average set size; $t_{\max}$, t_{avg}: maximum and average participation of a vertex to a set.

Dataset	$\|V\|$	$\|E\|$	$\|V_0\|$	$\|E_0\|$	ℓ	$\text{avg}(\alpha_0)$	$s_{\min}$	s_{avg}	$t_{\max}$	t_{avg}
DBLP	10001	27687	10001	22264	1767	0.58	6	7.46	10	1.31
Youtube	10002	72215	10001	15445	5323	0.69	2	4.02	82	2.14
KDD	2891	11208	1598	3322	5601	0.96	2	2.40	107	8.41
ICDM	3140	10689	1720	3135	5937	0.96	2	2.34	139	8.11
FB-circles	4039	88234	2888	55896	191	0.64	2	23.15	44	1.53
FB-features	4039	88234	2261	20522	1239	0.93	2	3.75	13	2.05
lastFM-artists	1892	12717	1018	2323	2820	0.89	2	2.91	221	8.08
lastFM-tags	1892	12717	855	1800	651	0.88	2	3.43	20	2.61
DB-bookmarks	1861	7664	932	1145	1288	0.97	2	2.27	27	3.13
DB-tags	1861	7664	1507	2752	4167	0.96	2	2.26	68	6.25

Table 2: Characteristics of edges selected as strong by *Greedy* and the two baselines. b: number of violated triangles in the solution divided by the number of open triangles (all possible violations); s: number of strong edges in the solution divided by the number of all edges; c: average number of connected components per community. A corresponds to *Angluin*; S corresponds to *Sintos*.

	Greedy			*Angluin*			*Sintos*		
Dataset	b	s	c	b_A/b	s_A/s	c_A	b_S/b	s_S/s	c_S
DBLP	0.07	0.47	1	2.77	0.77	1	0.0	1.08	3.53
Youtube	0.01	0.16	1	1.21	0.98	1	0.0	0.49	3.30
KDD	0.08	0.35	1	1.09	0.63	1	0.0	0.81	1.93
ICDM	0.07	0.38	1	1.06	0.57	1	0.0	0.83	1.84
FB-circles	0.002	0.15	1	61.05	0.20	1	0.0	1.05	8.76
FB-features	0.003	0.12	1	0.36	0.22	1	0.0	1.35	2.41
lastFM-artists	0.02	0.15	1	1.11	0.78	1	0.0	0.67	2.58
lastFM-tags	0.008	0.12	1	1.17	0.68	1	0.0	0.83	2.98
DB-bookmarks	0.01	0.35	1	1.01	0.35	1	0.0	1.04	1.61
DB-tags	0.10	0.45	1	1.02	0.66	1	0.0	0.80	1.74

Table 3: Precision and recall of *Angluin*.

Dataset	P_W	R_W	P_S	R_S
KDD	0.86	0.92	0.63	0.48
ICDM	0.87	0.93	0.66	0.50
lastFM-artists	0.91	0.95	0.54	0.37
lastFM-tags	0.92	0.95	0.26	0.16
DB-bookmarks	0.92	0.94	0.36	0.27
DB-tags	0.82	0.87	0.50	0.41

Table 4: Precision and recall of *Sintos*.

Dataset	P_W	R_W	P_S	R_S
KDD	0.78	0.70	0.19	0.26
ICDM	0.77	0.66	0.18	0.28
lastFM-artists	0.88	0.90	0.14	0.12
lastFM-tags	0.91	0.89	0.09	0.11
DB-bookmarks	0.92	0.64	0.13	0.49
DB-tags	0.75	0.62	0.22	0.35

and precision P_S and recall R_S for strong edges as

$$P_S = \frac{|S \cap E_{intra}|}{|S|} \quad \text{and} \quad R_S = \frac{|S \cap E_{intra}|}{|E_{intra}|}.$$

Angluin selects greedily edges that connect as many communities as possible. In other words, it prefers edges that are in many communities in the training set, and this acts as a strong signal for an edge being also in a community in the test set. The results shown in Tables 3–5 support this intuition, showing that *Angluin* obtains the best results. We also see that *Sintos*, which does not use any community information, has the worst results, while our method is able to improve *Sintos* by incorporating information from communities.

Running time. Our implementation was done in Python and the bottleneck of the algorithm is constructing the list of wedges. The running times vary greatly from dataset to dataset. At fastest we needed 52 seconds while the at slowest

Table 5: Precision and recall of *Greedy*.

Dataset	P_W	R_W	P_S	R_S
KDD	0.85	0.75	0.36	0.51
ICDM	0.85	0.71	0.34	0.55
lastFM-artists	0.91	0.90	0.36	0.39
lastFM-tags	0.92	0.90	0.15	0.20
DB-bookmarks	0.93	0.67	0.15	0.57
DB-tags	0.81	0.66	0.32	0.55

we needed almost 5 hours. We should point out that a more efficient implementation as well using parallelization with constructing the wedges should lead to significant reduction in computational time.

Case study. To demonstrate a simple use case, we use a snippet of *KDD* dataset: We picked five recent winners of SIGKDD innovation award: Philip S. Yu, Hans-Peter Kriegel, Pedro Domingos, Jon M. Kleinberg and Vipin Kumar and constructed an underlying network as a union of their ego-nets. We then used 5 common topics, *cluster*, *classif*, *pattern*, *network*, and *distribut* as communities. Figure 2 depicts the discovered edges of *Greedy*. From the figure we see that showing only strong edges significantly simplifies the graph. The selected strong edges are reasonable: for example a path from Hans-Peter Kriegel to Pedro Domingo was Arthur Zimek, Karsten Borgwardt, and Luc De Raedt, while a path from Pedro Domingo to Jon Kleinberg was Luc De Raedt, Xifeng Yan, Zhen Wen, Ching-Yung Lin, Hang-hang Tong, Spiros Papadimitriou Christos Faloutsos, and Jure Leskovec.

A zoom-in version of the graph of Figure 2, showing the names of all authors, is omitted due to space constraints but can be found in the public code and dataset repository.[6]

7 CONCLUDING REMARKS

We presented a novel approach for the problem of inferring the strength of social ties. We assume that social ties can be one of two types, strong or weak, and as a guiding principle for the inference process we use the strong triadic closure property. In contrast to most works that use interaction data between users, which are private and thus, typically not available, we also consider as input a collection of *tight* communities. Our assumption is that such tight communities are connected via strong ties. This assumption is valid in cases when being part of a community implies a strong connection to one of the existing members. For instance, in a scientific collaboration network, a student is introduced on a research topic by his/her supervisor who is already working in that topic.

Based on the STC principle and our assumption about community-level connectivity, we formulate two variants of the tie-strength inference problem: MinViol, where we ask to minimize the number of STC violations, and MaxTri, where we the goal is to maximize the number of non-violated open triangles. We show that both problems are **NP**-hard. Furthermore, we show that the MinViol problem is hard to approximate, while for MaxTri we develop an algorithm with approximation guarantee. For the approximation algorithm we use a greedy algorithm for maximizing a submodular function on intersection of matroids.

There are many interesting directions to explore in the future. An interesting question is to consider alternative problem formulations that combine the strong triadic closure property with other community-level constraints, such as density and small diameter. We would also like to consider formulations that incorporate user features. A different direction is to consider an interactive version of the problem, where the goal is to select a small number of edges to query, so that the correct labeling on those edges can be used to maximize the accuracy of inferring the strength of the remaining edges.

Acknowledgements. This work was supported by the Tekes project "Re:Know", the Academy of Finland project "Nestor" (286211), and the EC H2020 RIA project "SoBigData" (654024).

[6] https://github.com/polinapolina/connected-strong-triadic-closure

REFERENCES

[1] Dana Angluin, James Aspnes, and Lev Reyzin. 2013. Network construction with subgraph connectivity constraints. *Journal of Combinatorial Optimization* (Jan. 2013).
[2] Lars Backstrom and Jon Kleinberg. 2014. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 831–841.
[3] Ronald S Burt. 2009. *Structural holes: The social structure of competition*. Harvard university press.
[4] Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit S Dhillon, and Ambuj Tewari. 2014. Prediction and clustering in signed networks: a local to global perspective. *Journal of Machine Learning Research* 15, 1 (2014), 1177–1213.
[5] James Davis. 1970. Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *American Sociological Review* (1970), 843–851.
[6] J. Davis and S. Leinhardt. 1972. The structure of positive interpersonal relations in small groups. *Sociological Theories in Progress* (1972), 218–251.
[7] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2014. On Facebook, most ties are weak. *Commun. ACM* 57, 11 (2014), 78–84.
[8] David Easley and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
[9] Zhanpeng Fang and Jie Tang. 2015. Uncovering the Formation of Triadic Closure in Social Networks. In *IJCAI*. 2062–2068.
[10] Marshall L Fisher, George L Nemhauser, and Laurence A Wolsey. 1978. An analysis of approximations for maximizing submodular set functions-II. In *Polyhedral combinatorics*. Springer, 73–87.
[11] Eric Gilbert. 2012. Predicting tie strength in a new medium. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1047–1056.
[12] Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 211–220.
[13] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.
[14] Jacob Holm, Kristian De Lichtenberg, and Mikkel Thorup. 2001. Poly-logarithmic deterministic fully-dynamic algorithms for connectivity, minimum spanning tree, 2-edge, and biconnectivity. *J. ACM* 48, 4 (2001), 723–760.
[15] Jason Jones, Jaime Settle, Robert Bond, Christopher Fariss, Cameron Marlow, and James Fowler. 2013. Inferring tie strength from online directed behavior. *PloS one* 8, 1 (2013), e52168.
[16] Indika Kahanda and Jennifer Neville. 2009. Using Transactional Information to Predict Link Strength in Online Social Networks. *ICWSM* 9 (2009), 74–81.
[17] Gueorgi Kossinets and Duncan Watts. 2006. Empirical analysis of an evolving social network. *Science* 311, 5757 (2006), 88–90.
[18] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World Wide Web*. ACM, 641–650.
[19] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data. (June 2014).

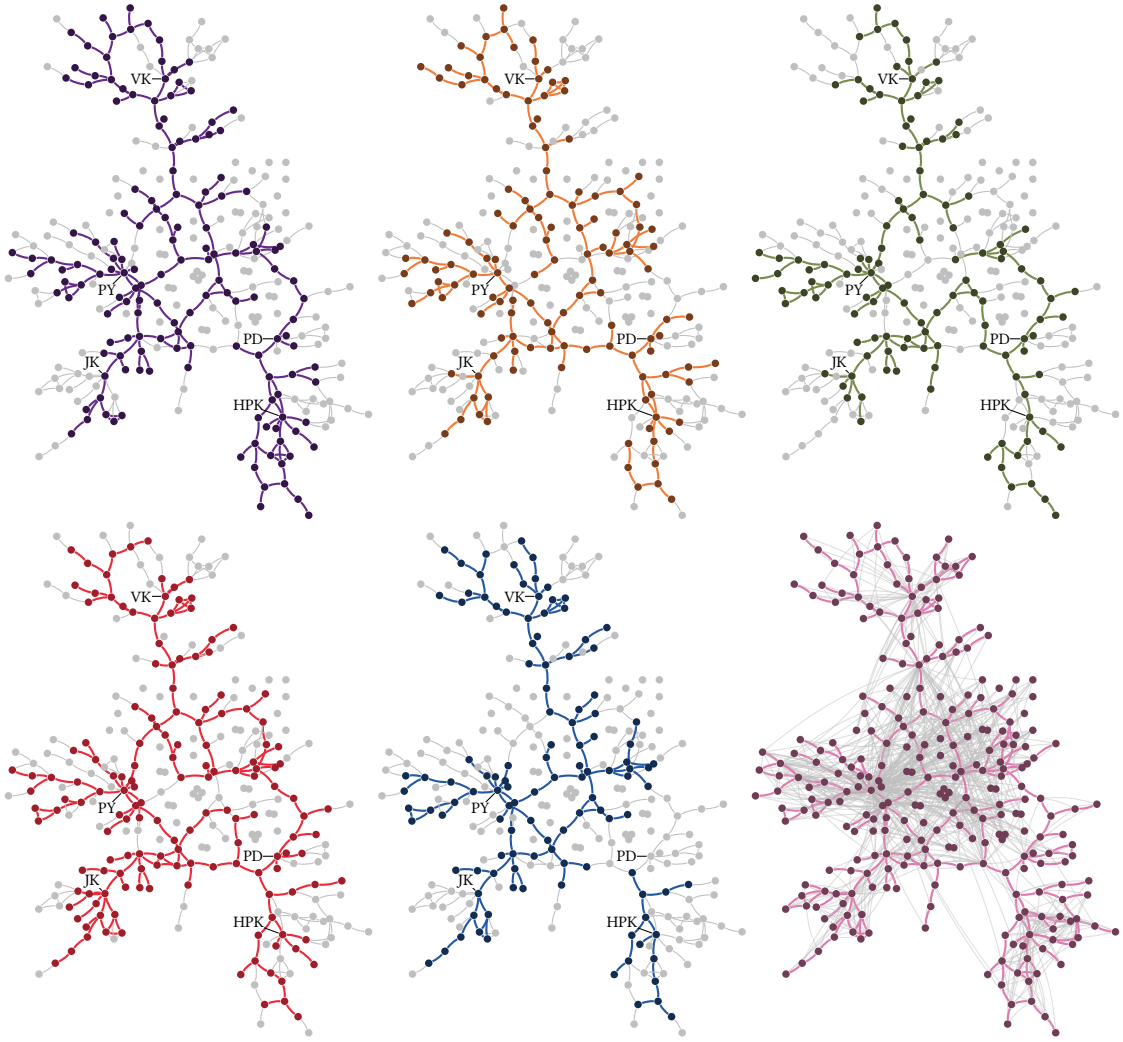

Figure 2: Discovered strong edges of 5 ego-networks of KDD innovation award winners. The first 5 figures contain only strong edges: the colored edges and vertices show 5 topics that were used as input: cluster, classif, pattern, network, distribut. The last topic consisted of 2 connected components which we used as two separated communities. The last figure shows strong *and* weak edges. Some of the vertices do not belong to any of the communities. Some edges are strong despite not belonging to any of the communities because we keep edges that do not induce violations.

[20] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.

[21] Linyuan Lü and Tao Zhou. 2010. Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)* 89, 1 (2010), 18001.

[22] Haris Memic. 2009. Testing the strength of weak ties theory in small educational social networking websites. In *International Conference on Information Technology Interfaces*. IEEE, 273–278.

[23] James D Montgomery. 1992. Job search and network composition: Implications of the strength-of-weak-ties hypothesis. *American Sociological Review* (1992), 586–596.

[24] T. M. Newcomb. 1961. *The acquaintance process*. Holt, Rinehart & Winston.

[25] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104, 18 (2007), 7332–7336.

[26] James Oxley. 2003. What is a matroid? *Cubo Matemática Educacional* 5, 3 (2003), 179–218.

[27] Stavros Sintos and Panayiotis Tsaparas. 2014. Using strong triadic closure to characterize ties in social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM, 1466–1475.

[28] Jie Tang, Tiancheng Lou, and Jon Kleinberg. 2012. Inferring social ties across heterogenous networks. In *Proceedings of the fifth ACM international conference on Web Search and Data Mining*. ACM, 743–752.

[29] Rongjing Xiang, Jennifer Neville, and Monica Rogati. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World Wide Web*. ACM, 981–990.

[30] Wayne W Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of anthropological research* 33, 4 (1977), 452–473.