MINING SOCIAL AND CRYPTOCURRENCY NETWORKS

PH.D. THESIS

BOOKLET

Author Ferenc Béres
Institute for Computer Science and Control
Eötvös Loránd Research Network

Supervisor András A. Benczúr
Institute for Computer Science and Control
Eötvös Loránd Research Network

DOI 10.15476/ELTE.2022.171

EÖTVÖS LORÁND UNIVERSITY

DOCTORAL SCHOOL OF INFORMATICS

2022

In the last decade, network science was flourishing since graph structures underlie several applications that we use during our daily routine. Social networks are probably the largest source for graph data. Facebook and Instagram both gained billions of new active users during this period[1]. Another important graph data source is related to mobility, flights between cities, ride-hailing, or route-planning applications. Finally, for most cryptocurrencies, there is an underlying transaction network where users exchange their coins or other funds without any governmental or third-party supervision. In contrast to most social network and user mobility platforms, the user interactions for various cryptocurrency networks are available to everyone due to the public nature of the blockchain. That is why Bitcoin and Ethereum, the two most well-known cryptocurrency networks, are also in the focus of my research.

A general problem with graph data is that it cannot be fed to classical machine learning methods in a straightforward way. Algorithms like logistic regression, decision trees or deep neural networks only work well with tabular data. Due to the irregular size of node neighborhoods, raw network data cannot be considered tabular. One possibility to solve general graph mining tasks such as node classification, link prediction, or community detection is to learn a vector space representation of network nodes for downstream machine learning tasks. Research in the related field of node embedding was recently catalyzed by the Word2Vec algorithm [8] for learning word representations in human language text. The main idea of network embedding is to explore the graph through multiple random walks and feed these node sequences to a neural network architecture (e.g. Skip-Gram model) that learns a representation for every node. The time complexity of this technique is linear in the number of vertices thus it can also be deployed for large networks with millions of nodes.

Unfortunately, linear time complexity in the number of nodes can be prohibitive for real-time dynamic network applications. In many data-intensive tasks where interactions between network participants are constantly arriving over time, we need to update graph mining models regularly to capture the latest changes in the data distribution, such as sudden bursts in popularity or some irregular network behavior. Fitting batch algorithms for large graph snapshots could cause a significant time-delay in the prediction. That is why online graph learning techniques are much preferred in these scenarios.

The main goal of our research is to analyze and model user behavior in social and cryptocurrency networks. Specifically, we intend to answer the following questions:

- What are the main advantages of online graph mining techniques over batch models for large-scale social networks and how to best compare their performance? In our research, we focus on graph centrality and node embedding techniques.
- How to mine cryptocurrency networks with novel network science tools to answer open questions in the domain of cryptoeconomics and privacy?

By collecting various new Twitter and cryptocurrency network data sets, we were among the first to deploy and analyze node embedding models in several network applications such as vaccine skepticism detection or Ethereum address deanonymization.

Our findings are related to the fields of network science and machine learning. In our work, we analyze user interactions in social and cryptocurrency networks as well as user-related metadata that we used to formulate supervised evaluation for most of the addressed graph mining tasks.

[1] https://www.businessofapps.com/data/facebook-statistics/, https://www.businessofapps.com/data/instagram-statistics/

Our contributions

Next, we explain our main results one by one. For each topic, we list our main contributions and the original source of publication.

1 Temporal networks

First, we introduce temporal networks by laying out the theoretical background for two dynamic graph computational models, the snapshot-based and the edge stream approaches. We rigorously compare these concepts by following the arguments in our recent tutorial:

[BBKP2021] András Benczúr, **Ferenc Béres**, Domokos Kelen, and Róbert Pálovics. Tutorial on graph stream analytics. DEBS '21, page 168–171, New York, NY, USA, 2021. Association for Computing Machinery.

In order to quantitatively analyze the performance of selected online graph algorithms, I collected two Twitter data sets, *RG17* and *UO17*, related to Roland-Garros 2017, the French Open Tennis Tournament, and to US Open 2017, the United States Open Tennis Championship. For both of these sport events, I collected tweets containing predefined hashtags, and then I extracted the underlying @-mention network. The following properties make *RG17* and *UO17* highly suitable for evaluating algorithms on dynamic graphs:

- Temporal @-mention network: every @-mention link in the graph has a timestamp covering a long time range.
- Large scale: both mention graph contains more than 300K edges and 70K nodes (Twitter accounts).
- Dynamic temporal ground truth information available from an external source, which makes supervised evaluation possible for edge stream based online graph algorithms. Based on the official event schedule, I compiled a binary node relevance label with daily granularity for the nodes of the *RG17* and *UO17* mention networks.

We first described the *RG17* and *UO17* Twitter collections in

[BPOB2018] **Ferenc Béres**, Róbert Pálovics, Anna Oláh, and András A Benczúr. Temporal walk based centrality metric for graph streams. *Applied Network Science*, 3(32):26, 2018.

Both of these data sets were included in a publication that received the best resource paper award at CIKM '21,

[R+2021] Benedek Rozemberczki, Paul Scherer, Yixuan He, George Panagopoulos, Alexander Riedel, Maria Astefanoaei, Oliver Kiss, **Ferenc Béres**, Guzmán López, Nicolas Collignon, and Rik Sarkar. Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models. CIKM '21, page 4564–4573, New York, NY, USA, 2021. Association for Computing Machinery.

2 Temporal walk based centrality metric for graph streams

The definitions of centrality can vary greatly and can incorporate both global and local factors of a user's location within the social network [2]. For temporal networks, a few generalizations of static centrality measures to dynamic settings have been suggested recently [1, 5, 7, 10, 11]. In

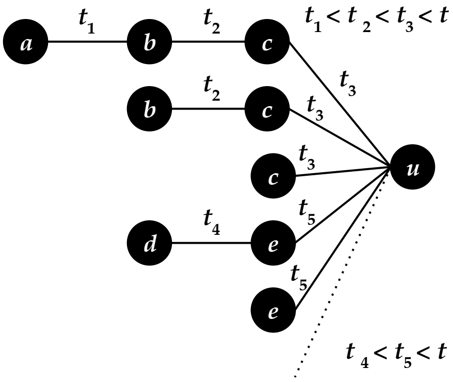

Figure 1: Definition of temporal Katz centrality: weighted sum of time-respecting walks ending at node u up to time t.

these works, tracking centrality of a single node and determining its variability play a major role [11], as it has been observed in the literature that centrality of nodes can change drastically from one time period to another [3].

We addressed a practically important variant of dynamic centrality: Our goal was to compute online updateable measures that can be computed from a data stream of time-stamped edges. The above results [1,5,7,10,11], however, cannot be used for computing and updating centrality online. To the best of our knowledge, only two previous studies [4, 9] propose data stream updateable centrality measures. Our approach is similar to [9] that also uses the notion of time-respecting walks to extend PageRank to the edge stream dynamic graph model.

Our contributions:

Thesis 1: *We defined temporal Katz centrality, an online updateable centrality metric based on time-respecting walks, as illustrated in Figure 1.*

- We defined the Temporal Katz centrality of node u at a given time t as the weighted sum of all time-respecting walks that end in u up to time t.
- We incorporated arbitrary time decay functions in our Temporal Katz centrality measure that can be adapted to the task in question.
- We gave online update algorithms for Temporal Katz, making it ideal for data-intensive applications.
- We gave two convergence theorems that mathematically justify the connection between our method and Katz index [6].
- We conducted a supervised evaluation on our Twitter data collections, *RG17* and *UO17*, introduced in the previous chapter. Using only network centrality, we tried to detect daily tennis player accounts as early as possible. Our measurements on these data sets show that temporal Katz centrality outperforms both static and online baselines.
- Finally, we performed extensive parameter analysis for properties such as score variability between consecutive snapshots as well as adaptation to concept drift.

Our results are published in

[BPOB2018] **Ferenc Béres**, Róbert Pálovics, Anna Oláh, and András A Benczúr. Temporal walk based centrality metric for graph streams. *Applied Network Science*, 3(32):26, 2018.

3 Node embeddings in dynamic graphs

Next, I investigated methods to encode (*embed*) the nodes of a dynamic network by vectors in a low-dimensional vector space in a way that representations in the embedded space reflect the neighborhood or structural properties of the nodes in the original graph.

Over the last years, a myriad of static node embedding methods have been proposed and applied in node classification and link prediction tasks. In our work, we propose two data stream updateable node embedding methods, StreamWalk and Online Second Order Similarity, with an application similar to static embedding models.

Our contributions:

Thesis 2: *We proposed two online updateable node embedding algorithms, StreamWalk and Online Second Order Similarity, that are both able to efficiently maintain representations of network nodes as the graph evolves over time.*

- We described StreamWalk, an online node embedding algorithm. Similar to temporal Katz centrality [BPOB2018], StreamWalk is also based on time-respecting walks.
- We described Online Second Order Similarity, which directly learns the neighborhood similarity of node pairs in the graph stream by approximating their neighborhood Jaccard similarity at a given time.
- We conducted supervised node similarity search evaluation on our *RG17* and *UO17* Twitter data collections. We showed that our models can efficiently differentiate daily tennis player accounts from other network participants. Our measurements on these data sets show that our online node embedding models outperform static baselines such as LINE, node2vec or DeepWalk.
- Finally, we showed that the combination of StreamWalk and Online Second Order Similarity further improves the accuracy of similarity search.

Our results are published in

[BKPB2019] **Ferenc Béres**, Domokos M. Kelen, Róbert Pálovics, and András A Benczúr. Node embeddings in dynamic graphs. *Applied Network Science*, 4(64):25, 2019.

4 Vaccine skepticism detection by network embedding

As an application, we deploy node embedding models for vaccine skepticism detection. We analyze social network data related to Covid-19 vaccination. We focus on two groups of people commonly referred to as pro-vaxxers and vax-skeptic users. In short, the first group supports vaccination, while the second questions vaccine efficacy or the need for general vaccination against Covid-19. We intended to develop techniques that can efficiently differentiate content based on the expressed vaccine view.

Our contributions are the following:

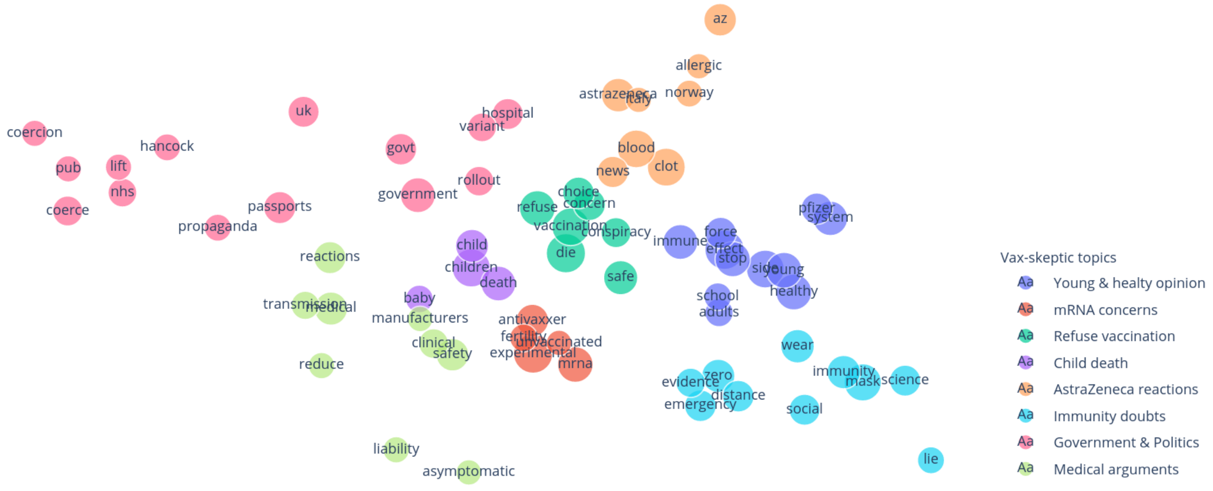

Figure 2: Vax-skeptic topic space uncovered by node embeddings. In the center, there are anti-vaxxer topics (e.g. child death cases, fear from the mRNA technology) that are surrounded by less offensive discussions (e.g. politics, medical arguments, immunity concerns).

Thesis 3: *By learning meaningful information from the Twitter reply network, node embedding models can efficiently detect vaccine skepticism in online content (tweets).*

- We collected and annotated a large Twitter data set related to Covid-19 vaccination.
- We quantitatively assessed the performance of node embedding for the task of vaccine skepticism detection by deploying them on the reply network that we extracted from the data.
- By training a binary classifier to predict the expressed vaccine view for each tweet, we found that node embedding models can significantly improve performance compared to text-only approaches. Furthermore, they can even reveal pro-vaxxer and vax-skeptic user clusters as well as their underlying topic hierarchy, see Figure 2.
- We released our data and source code on GitHub.

We presented our results at a conference:

[BCMB2021] Ferenc Béres, Rita Csoma, Tamás Vilmos Michaletzky, and András A. Benczúr. Vaccine skepticism detection by network embedding. In *Book of Abstracts of the 10th International Conference on Complex Networks and Their Applications*, pages 241–243, 2021.

5 Profiling and Deanonymizing Ethereum Users

Ethereum is the largest public blockchain by usage. It is an account-based cryptocurrency where users store their assets in accounts that they tend to frequently re-use to interact with a wide range of services and decentralized applications (e.g. games, exchanges). As it is a blockchain-based cryptocurrency, the transaction history for each account is publicly observable.

In our experiment, I embedded the nodes of the Ethereum transaction graph to profile and deanonymize Ethereum users based on their network activity. The nodes in this graph are

Ethereum addresses (accounts) and the transactions are directed links between them. Each physical entity (e.g. users, companies) may own multiple addresses, and the exact address-entity relations are usually hidden from the public. Thus, I rigorously analyze the Ethereum transaction network to reveal these connections. In the cryptocurrency domain, we were the first to quantitatively assess the performance of a recent area of machine learning in graphs, the so-called node embedding algorithms.

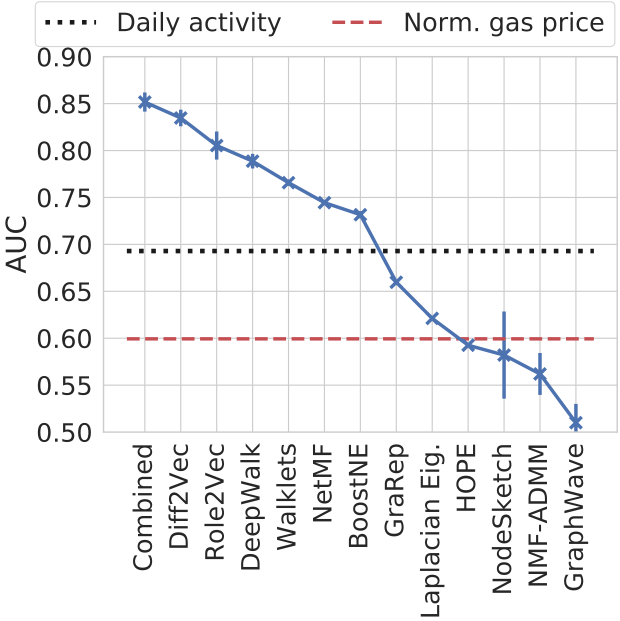

Figure 3: Deanonymization task: find accounts of the same user. AUC is presented for 13 node embedding models as well as time-of-day activity and gas price profile based baselines (horizontal lines).

Our contributions:

Thesis 4: *Node embedding models can efficiently link addresses that belong to the same user.*

- We collected Ethereum related data from several sources, including Ethereum name service (ENS), Etherscan blockchain explorer, Tornado Cash mixer contracts, and Twitter.
- Using ENS identifiers as ground truth information, we quantitatively compared multiple node embedding models in a deanonymization task where we link accounts of the same user. As illustrated in Figure 3, some node embedding methods significantly outperform user activity based baselines.
- As a direct application, we showed that node embedding based profiling can significantly decrease the privacy guarantees of the Tornado Cash (TC) mixer service, which was originally proposed to obfuscate the relationship between addresses of the same user.
- Finally, in light of our results, we proposed a few best practices for Ethereum users to follow in order to increase their privacy.

Our results appeared in

[BSBQ2021] **Ferenc Béres**, István András Seres, András A Benczúr, and Mikerah Quintyne-Collins. Blockchain is watching you: Profiling and deanonymizing ethereum users.

In *2021 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*, pages 69–78, 2021.

6 Cryptoeconomic traffic analysis of Bitcoin's Lightning network

Finally, we analyzed the Lightning Network (LN), a payment channel network that was designed to solve Bitcoin's scalability issues. It allows participants to exchange transactions locally, without broadcasting them to the blockchain. Thus, LN opens the way for instant low-value payments with negligible fees.

In a payment channel network, nodes are users and the edges are payment channels. A given node can issue payments only to those participants that it can reach through a series of edges. Intermediary nodes of a given payment path can independently decide the transaction fees that they charge for relaying the payment.

Our contributions:

As original LN payments are cryptographically hidden from us, we designed a payment traffic simulator to quantitatively confirm several concerns related to LN that the cryptocurrency community had been speculating about for a long time. A main contribution compared to previous simulation-based studies was that we managed to identify more than 100 merchant nodes on LN.

Thesis 5: *By simulating LN payments from ordinary users towards merchants, we found that central router nodes have (1) low annual RoI, and (2) strong statistical evidence on payment sender and receiver nodes.*

By simulating payments at different value and daily transaction volume levels, we made several observations related to the state of LN in 2019:

- We concluded that low routing fees do not sufficiently compensate the routing nodes that essentially hold the network together. Based on our measurements, the annual return of investment (RoI) for every major router is less than 4%. However, they could achieve

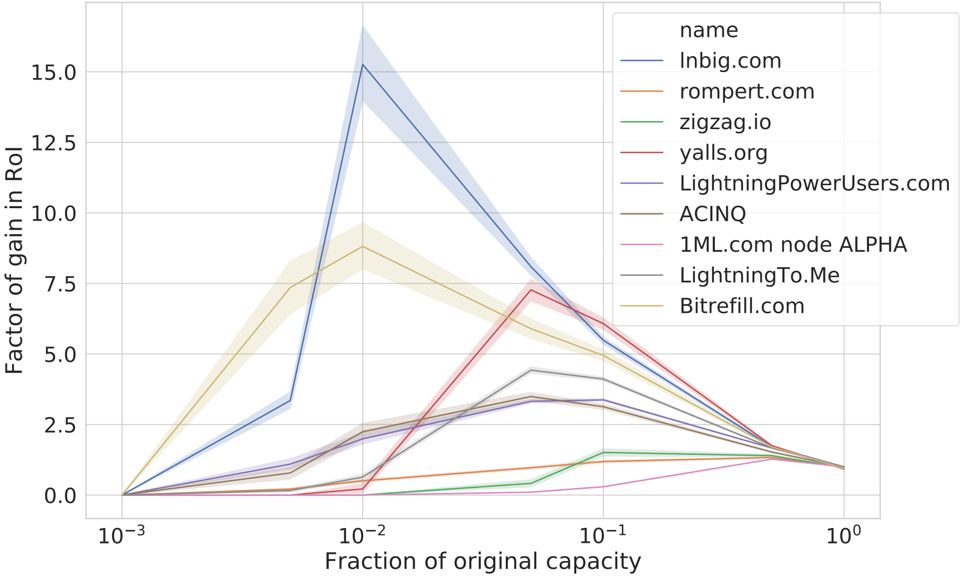

Figure 4: RoI gain after reducing node capacities to the given fractions.

significantly better RoI, shown in Figure 4, by reducing capacity on their currently over-provisioned payment channels.

- We further assessed the importance of router entities by monitoring the changes in the number of failed payments after we exclude them one by one from LN.
- Finally, we observed that despite onion routing, routers can gather strong statistical evidence about the sender and receiver of LN payments, since a substantial portion of payments involves only a single routing intermediary. Thus, we propose to use longer, suboptimal paths to gain more privacy. Our genetic algorithm based solution only marginally increases the costs for LN users.

Our results were published in

[BSB2021] **Ferenc Béres**, István András Seres, and András A Benczúr. A cryptoeconomic traffic analysis of bitcoin's lightning network. *Cryptoeconomic Systems*, 1(1), 2021.

7 Credits

The first part of my research is related to temporally evolving networks. We developed new online network centrality and node embedding techniques that outperformed existing snapshot-based approaches. In this work, I collected and annotated dynamic network data, implemented and measured most of the algorithms. Róbert Pálovics and Domokos Miklós Kelen participated in node embedding model implementations [BKPB2019]. They also verified experimental results and contributed to algorithm descriptions in our articles [BPOB2018, BKPB2019].

My research related to cryptocurrency networks published in [BSB2021,BSBQ2021] is joint work with István András Seres, who contributed with his knowledge on cryptocurrencies, defined the problems, and described the cryptocurrency related background in both papers. The analysis of the basic graph properties of the Bitcoin Lightning Network and their change in time [BSB2021] is also his contribution. In our works, I designed, implemented, and evaluated the experiments related to traffic simulation and node embedding. Finally, I augmented and collected Bitcoin and Ethereum related cryptocurrency network data sets that are rigorously assessed in my Thesis.

My Publications

[BBKP2021] András Benczúr, Ferenc Béres, Domokos Kelen, and Róbert Pálovics. Tutorial on graph stream analytics. DEBS '21, page 168–171, New York, NY, USA, 2021. Association for Computing Machinery.

[BCMB2021] Ferenc Béres, Rita Csoma, Tamás Vilmos Michaletzky, and András A. Benczúr. Vaccine skepticism detection by network embedding. In *Book of Abstracts of the 10th International Conference on Complex Networks and Their Applications*, pages 241–243, 2021.

[BKPB2019] Ferenc Béres, Domokos M. Kelen, Róbert Pálovics, and András A Benczúr. Node embeddings in dynamic graphs. *Applied Network Science*, 4(64):25, 2019.

[BPOB2018] Ferenc Béres, Róbert Pálovics, Anna Oláh, and András A Benczúr. Temporal walk based centrality metric for graph streams. *Applied Network Science*, 3(32):26, 2018.

[BSB2021] Ferenc Béres, István András Seres, and András A Benczúr. A cryptoeconomic traffic analysis of bitcoin's lightning network. *Cryptoeconomic Systems*, 1(1), 2021.

[BSBQ2021] Ferenc Béres, István András Seres, András A Benczúr, and Mikerah Quintyne-Collins. Blockchain is watching you: Profiling and deanonymizing ethereum users. In *2021 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*, pages 69–78, 2021.

[R+2021] Benedek Rozemberczki, Paul Scherer, Yixuan He, George Panagopoulos, Alexander Riedel, Maria Astefanoaei, Oliver Kiss, Ferenc Béres, Guzmán López, Nicolas Collignon, and Rik Sarkar. Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models. CIKM '21, page 4564–4573, New York, NY, USA, 2021. Association for Computing Machinery.

References

[1] Ahmad Alsayed and Desmond J Higham. Betweenness in time dependent networks. *Chaos, Solitons & Fractals*, 72:35–48, 2015.

[2] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262, 2014.

[3] Dan Braha and Yaneer Bar-Yam. From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity*, 12(2):59–63, 2006.

[4] Marwan Ghanem, Florent Coriat, and Lionel Tabourier. Ego-betweenness centrality in link streams. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 667–674, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Peter Grindrod and Desmond J Higham. A dynamical systems view of network centrality. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 470, 2014.

[6] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[7] Hyoungshick Kim and Ross Anderson. Temporal node centrality in complex networks. *Physical Review E*, 85(2):026107, 2012.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] Polina Rozenshtein and Aristides Gionis. Temporal pagerank. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 674–689. Springer, 2016.

[10] John Tang, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Vincenzo Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, page 3. ACM, 2010.

[11] Dane Taylor, Sean A Myers, Aaron Clauset, Mason A Porter, and Peter J Mucha. Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574, 2017.