EventOrient: A System for Community-Scale Tempo-Contextual Network Analysis of Social Data Streams

Ammar Raşid Abdullah Ihsan Seçer Abdurrahman Aboudakika Ahmet Öztemiz
Mustafa Bera Akay
Istanbul Şehir University
34865 Dragos, Istanbul, Turkey
{ammarrashed,abdullahsecer,talaataboudakika,ahmetoztemiz,mustafaakay}@std.sehir.edu.tr

July 6, 2018

Abstract

We study the evolution of the social network of Istanbul Şehir University overtime, capturing direct, contextual and latent changes in the network structure. The university's story is embodied in the networks we construct, analyze and temporally monitor. Our system,*EventOrient*, stands on three components; Web Crawling, Networked Data Analysis and Data storytelling, making it a comprehensive system for community-scale tempo-contextual network analysis. Our goal is to render the social development of the university's community in a lucid and insightful manner.

1 Introduction

The relational structure of fully contextual social data is a loud yet subtle story of the society it represents. However, obtaining the full context of social data has proven to be highly infeasible. This infeasibility is both conceptual and computational. The conceptual infeasibility is associated with the lack of means to accurately and thoroughly measure the subjective characteristics of social contexts. The computational infeasibility is associated with the massive size of social data and the lack of sufficient computational capacity to store and process the varied and many contexts of social data.

This infeasibility is refuted by focusing on more specific communities. The specificity of a community may be geographic, such as a town or a region. A community may also be specified by an affiliation, such as an organization, a religion or a language. Time can also be a powerful dimension in classifying communities. The goal of specifying a community is to capture as much of meaningful contexts pertaining to the intrinsic structure of that community while maintaining reasonable feasibility.

EventOrient focuses on the community of Istanbul Şehir University (ISU). In such an affiliation-specific community it is important to capture *sub-affiliations*, e.g. student clubs, within the network to capture deeper insights. We classify complementary contexts horizontally, as mutable and immutable. A mutable contexts for a node in the social network can be fields of study, jobs or residence addresses. An immutable context can be native language, gender or birth date. Complementary contexts are classified vertically, as direct and implicit. Direct contexts are objective information that account directly to a structural value in the social network, such as *follows* on Twitter or Facebook's groups membership. Whereas implicit contexts are inferred from processed social data, such as topics of interest, general sentiments about particular facts and events or style of writing. The implicitness of a context does not necessarily imply more significance or relevance than direct contexts.

1.1 Motivation

Social computing systems are essential for providing crucial insights for social studies, organizational intelligence and crisis evasion. In this paper, we present *EventOrient*, a system for mining, analyzing and monitoring social data within Istanbul Şehir University's community.

2 Methodology

2.1 System Overview

EventOrient provides a thorough pipeline for social network analysis. This pipeline consists of three main components; *Web Crawling*, *Networked Data Analysis* and *Data Storytelling*.

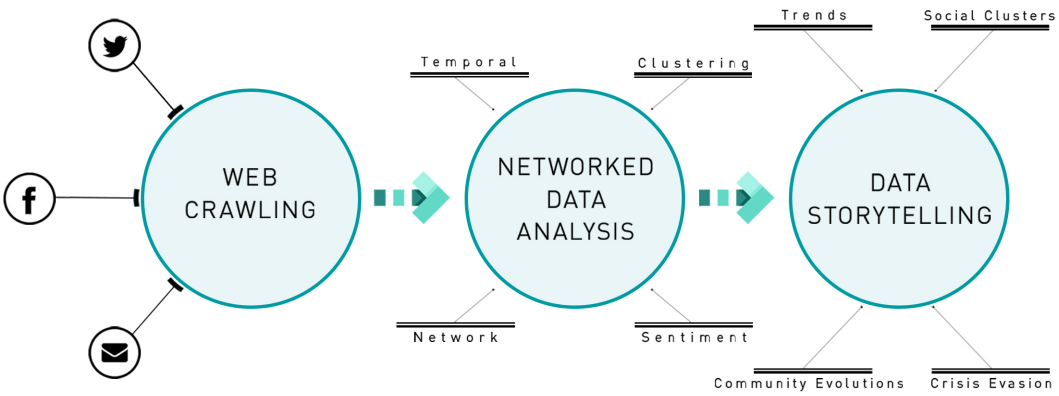


Figure 1: *EventOrient* System workflow

2.2 Web Crawling

Using snowball sampling technique, we iteratively crawl twitter users and their connections (follower and followee) with a custom depth level. We use *gmail* contacts within the domain of the university to match twitter accounts' names against for filtering. The matching ratio is a zero-to-hundred similarity score based on levenshtein distance [1] and is reported in the database to allow filtering out users with matching ratio less than a custom threshold t. The starting point of the crawler is institutional accounts. Institutional accounts enables crawling newcomers to the community in subsequent runs. After the crawler's first run, twitter accounts matching a name in the gmail contacts pool with a ratio higher than the threshold are used as starting point for the crawler in the subsequent runs. A connection is either present or absent. The database reports the date of the addition or the removal of a connection only if the previous state was different. Although this is a violation for the first normal form (1NF) [2], it has proved to be more efficient than decomposing the connections table and adding a new column for the date of each new run.

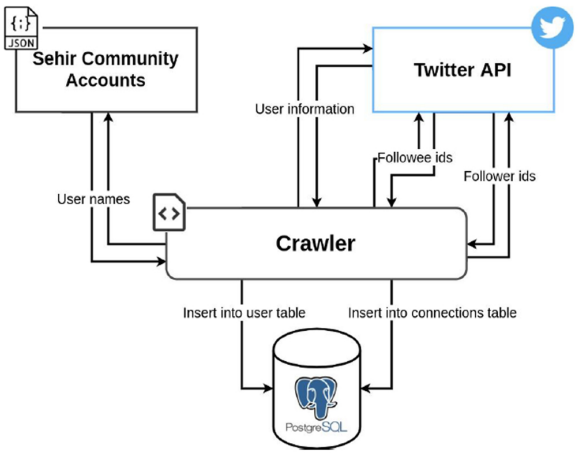


Figure 2: Web Crawling overview

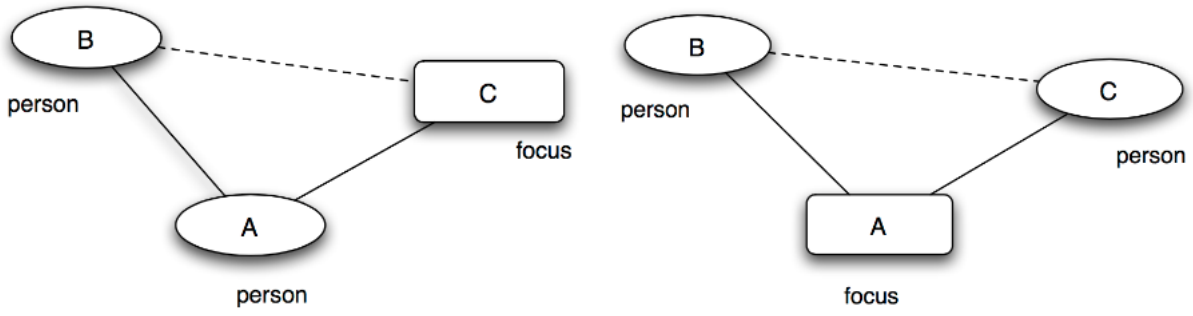
2.3 Networked Data Analysis

2.3.1 Static Network Analysis

Setting threshold $t = 90\%$, we construct a social network from filtered twitter accounts and their connections. We used NetworkX [3] to calculate nodes' degrees, eigenvector centralities [4], closeness centralities [5], betwenness centralities [6], clustering coefficients [7] and pagerank [8] scores. We use Girvan-Newman [9] algorithm to detect communities within the network. For institutional accounts, we label them as *foci*. Modularity [9], transitivity and network diameter are also calculated and reported. *Homophily* [10] is the tendency of people to form bonds with similar others. It has the potential of revealing clique-like behaviour and segregation in a network. We chose language to be the characteristic by which we measure *homophily*, as language proved the best accessibility to contextual relevance ratio.

2.3.2 Temporal Network Analysis

Institutional accounts, labeled as *foci*, are used to construct the affiliation network of the community. Social influence is an implicit subjective character of social networks. However, temporally monitoring affiliation networks provide a reasonable tool to measure social influence. Namely, closures reflect the flow the social influence within the network. We classified edges in the network as user-user edges or focal edges. An edge $\{u, v\}$ is a focal edge if exactly one of the nodes u and v is labeled as *foci*. A focal edge $\{u, f\}$ between a user u and a foci f at time t_j completes a membership closure if there exists at time t_i at least one edge $\{u, v\}$ between u and another user v and a focal edge between v and f, where $t_i < t_j$. An edge $\{u, v\}$ between a user u and another



(a) Membership Closure

(b) Focal Closure

Figure 3: *Social Influence* is associated with membership closures, whereas *Selection* is associated with focal closures

user v at time t_j completes a focal closure if there exist at time t_i at least a *foci*-labeled node f with an edges $\{u, f\}$ and $\{v, f\}$. Membership closures reflect social influence; the tendency of people to modify their behaviour to align better with their friends. Whereas focal closures reflect selection; the tendency of people to form bonds with others sharing similar characteristic, usually immutable characteristic.

2.4 Data Storytelling

Django [11] framework is used to put together a dynamic interactive visualization. Networks constructed by the Analyzer are passed to *Django* templates as JavaScript Object Notation JSON responses. *Django*'s *Views* Functions filter the network as specified by the user interaction with the filtering widgets. Data-Driven Documents (D3) [12] library is used for network rendering.
Girvan-Newman [9] algorithm is used to detect communities within the network and label *user* nodes with their corresponding detected communities. Nodes with the same community labels are depicted with the same colors. Users can filter the network by the node-level metrics added by the Networked Data Analysis component. Sizes of the nodes are defined by a node-level metric customizable by the user. Users can filter nodes by their degrees, betweenness, closeness and eigenvector centralities, pagerank scores and clustering coefficients. Users can also choose which metric, of the aforementioned metrics, to be used as to scale nodes' sizes, in addition to

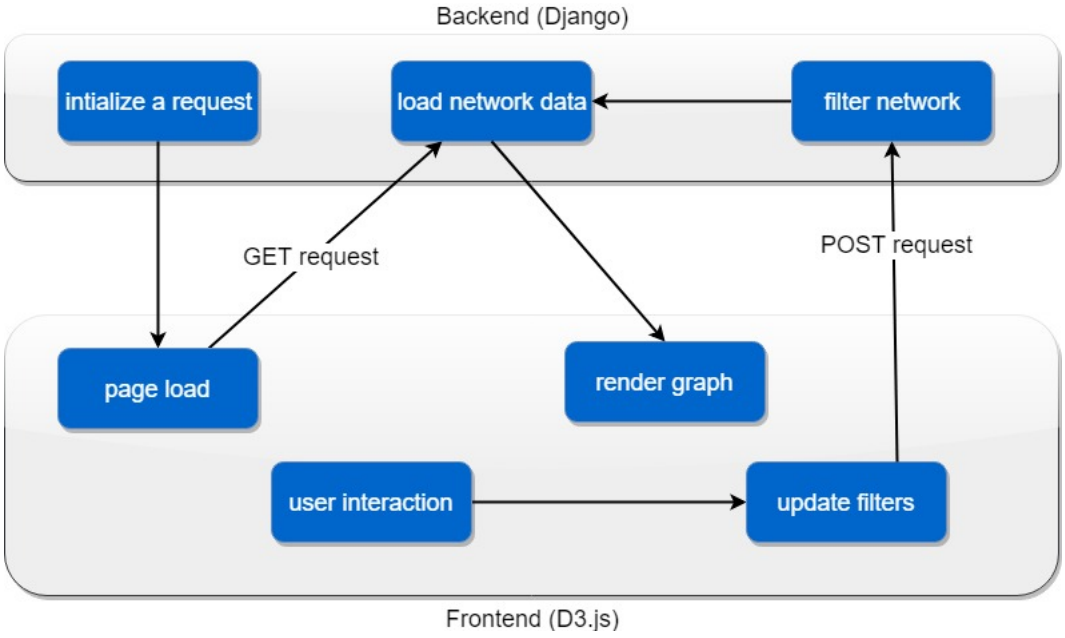


Figure 4: Interactive Visualization frontend-backend communication

in-degree, out-degree and followers count. For temporal filtering, a scroll is used to access the state of the network at different dates as it is being augmented regularly by the web crawling component. In order to ease searching for specific nodes in the network, a search bar provides searching by twitter's screen name, twitter's account names or gmail contact names. The set of nodes selected by the searching query are highlighted with higher opacity and bigger sizes, whereas other nodes' opacity is lowered. The graph automatically zooms on one of the selected nodes by the searching query, if any.

3 Findings and Discussion

3.1 Network Structure-related insights

According to the network state on $24/05/2018$, the following metrics were calculated:

	Affiliation Netw.	Users Netw.
Nodes	1041	554
Edges	2286	951
Avg. Degree	4.39	3.43
Betweenness	0.00066	0.00096
Pagerank	0.00096	0.0018
Closeness Cent.	0.0912	0.0617
Eigenvector Cent.	0.0087	0.014
Clustering Coeff.	0.306	0.113
Transitivity	0.0514	0.0455
Modularity	0.4535	0.619
Diameter	7	8

Table 1: Network state on 24/05/2018

3.2 Context-related insights

From $08/05/2018$ to $24/05/2018$ we found that there is a total of 638 new focal closures. Upon measuring *homophily* by language, we define a cross-metric edge as an edge $\{u, v\}$ between where u and v use different languages on their social media accounts. Cross-metric edge ratio (CMER) is the ratio of cross-metric edges to all edges. Heterogeneity Fraction Norm (HFN) is the probability of assigning a cross-metric edge in a randomly constructed network with the same distribution of nodes in the original network. We found that the probability of a link between two users speaking different languages in the network is higher than the heterogeneity fraction norm. That can be considered as an evidence for multi-cultural ambience in *Şehir* and refute the argument that language barrier in *Şehir*'s community is posing a significant hindrance to cross-cultural integration.

	Heterogeneity	Threshold
Affiliation Netw.	0.48250	0.38277
Users Netw.	0.51735	0.39109

Table 2: *Homophily* in the network

4 Conclusions

Thorough social computing systems are essential to render the story of a community in an intelligible and insightful manner. We have shown the feasibility and potential of studying affiliation-specific communities. Utilizing the right contextual metrics, we captured valuable subjective characteristics of *Şehir*'s social network. Introducing time and focal dimensions to our networked data analysis exposed significant implicit phenomena, such as *Selection* and *Social Influence*. *EventOrient* is adequately equipped for community-scale tempo-contextual networked data analysis, rendering the story of a community through myriad of direct, implicit, objective and subjective characteristics.

References

[1] F. P. Miller, A. F. Vandome, and J. McBrewster, *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press, 2009.

[2] M. A. Roth and H. F. Korth, "The design of 1nf relational databases into nested normal form," *SIGMOD Rec.*, vol. 16, pp. 143–159, Dec. 1987.

[3] D. A. Schult, "Exploring network structure, dynamics, and function using networkx," in *In Proceedings of the 7th Python in Science Conference (SciPy*, pp. 11–15, 2008.

[4] D. Taylor, S. A. Myers, A. Clauset, M. A. Porter, and P. J. Mucha, "Eigenvector-based centrality measures for temporal networks," *Multiscale Modeling & Simulation*, vol. 15, pp. 537–574, jan 2017.

[5] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, pp. 581–603, Dec 1966.

[6] L. Leydesdorff, "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 9, pp. 1303–1319, 2007.

[7] T. Opsahl, "Triadic closure in two-mode networks: Redefining the global and local clustering coeffi-

cients.," *Social Networks*, vol. 35, no. 2, pp. 159–167, 2013.

[8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," 1999.

[9] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[10] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," *ACM Trans. Web*, vol. 6, pp. 9:1–9:33, June 2012.

[11] C. Burch, "Django, a web framework using python: Tutorial presentation," *J. Comput. Sci. Coll.*, vol. 25, pp. 154–155, May 2010.

[12] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, pp. 2301–2309, Dec. 2011.