Orienting Social Event Streams as Data Stories

Conference Paper · December 2018

DOI: 10.1007/978-3-030-05918-7_13

4 authors, including:

Ammar Rashed
University of Ottawa
10 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Abdelrahman Aboudakika
Istanbul Sehir University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Ahmet Bulut
Giresun University
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Orienting Social Event Streams as Data Stories

Ammar Rashed, Abdullah İhsan Seçer,
Abdurrahman Aboudakika, and Ahmet Bulut

İstanbul Şehir University 34865 Dragos, İstanbul, Turkey
{ammarrashed,abdullahsecer,talaataboudakika}@std.sehir.edu.tr,
ahmetbulut@sehir.edu.tr

Abstract. We study the evolution of our university's social networks over time, capturing direct, contextual, and latent changes in these networks. With the assumption of our university's social dynamics being embodied in the networks we construct, we continuously monitor these networks in order to gain an understanding of the changes they go through and their evolution. Our system has three main components: (i) crawling the web for collecting data, (ii) networked data analysis, and (iii) data storytelling. Our goal is to render the social development of our university as a community in a lucid and insightful manner.

Keywords: Social Network Analysis · Data Science · Data Visualization.

1 Introduction

The digital ecosystem has experienced an exponential growth with billions of people and machines creating online content. Fueled by the proliferation of e-commerce and social media, the digital data revolution has necessitated the understanding of ever growing and large volumes of heterogeneous data for extracting key business insights and for gaining competitive advantage in order to survive in the age of data. Many institutions face an existential threat lest they fail to adapt to the impending data revolution. Irrespective of the scale of the institution and its business sector, data driven decision-making has become the heartbeat of business operations.

The data revolution has attracted a significant body of researchers to work on different aspects of social media from e-commerce to sociology, and to computer science. In this work, we focus on designing an intelligent system for social media. Our main goal is to create living and interactive stories off of the social media traces of any online community. The collected data is projected on to a live dashboard for monitoring the heartbeat of the community continuously. The dashboard is expected to expose key insights such as what excites the individual members of the community and who collectively moves them. Wu et al. built a system for visually exploring in multiple facets the digital footprint of events with high impact using social media streams [10]. The relationships between seemingly unrelated social media events were studied by Lu et al. in order to

answer questions such as whether positive tweets about a product proceed positive reviews about that product [6]. However, these recent studies mostly focus on synthesizing the media content and do not address how to visualize the social co-evolution of a network of individuals.

The structure of a network with its nodes and edges between pairs of nodes is a valuable conduit that enables us to study various population dynamics from how quickly new ideas spread through the network to where these ideas are adopted first if at all. Additionally, we can use the social media data of the users in the network to learn more about their interests and their general sentiment towards particular events that affect the community.

We identify and validate the members of a given community using its official social media accounts. Then, we periodically crawl the social media data of those members from select media outlets and ingest them into the backend datastore for network analysis. We identify micro communities in the community network and then study how they evolve over time, possibly changing the overall character of the macro or the parent community. We can further inspect the parent network and its micro communities in terms of their capacity for introversion (or extraversion) from coarse to finer grain using multi-modal data.

Note that the rendering and the encapsulation of the digital context of a given online community may be computationally infeasible simply because there are more things that we do not measure compared to what we actually measure about that community. This is partly due to the lack of means to accurately and thoroughly measure the objective as well as the subjective characteristics of a social context. Because of the inherent lack of data, any computational model we build may not fully represent the society we aim to model. However, even if each subtle yet possible data point may not be fully captured, we might still draw meaningful conclusions from the aggregate behavior of the individual members of the society and shed light onto how a certain call for change spawns and diffuses into the network to finally become visible in the aggregate. Therefore, we still resort to capturing as much data as possible that pertain to the intrinsic workings and the societal fabric of that community.

In the rest of the paper, we present the technical details of our system, which is currently under development. While developing the core parts of the system, we focused on our university and its community, i.e., our students, academics, and administrative personnel. We are designing it for monitoring the social media event streams of our university in order to orient its life-changing events as vivid data stories to tell. Currently, our data storytelling capacity is primitive: it corresponds to the visualization of the network and the synthesis of a case study we conducted on the network.

2 Methodology

2.1 Definitions

A network is a graph $G = (U, V)$ defined on a set of nodes V and a set of edges E. For two nodes $u, v \in V$, if there is an edge $e \in E$ between them, it is

denoted as $e = \{u, v\}$ for an undirected graph, and as $e = (u, v)$ for a directed graph. If the nodes of a given network are users exclusively, then it is called a social network. If the nodes are affiliations only, then it is called an affiliation network. However, if the nodes can be users as well as their affiliations, e.g., academic departments and social clubs, then this richer network is called a social-affiliation network [3]. Each edge in a social-affiliation network corresponds to either a user-to-affiliation relationship or a user-to-user relationship: on Twitter, a user can follow another user, and on Facebook, a user can join a group. In such affiliation-rich communities as universities, it is important to capture a diverse set of affiliations in order to gain deeper insights into the intrinsic workings of the community.

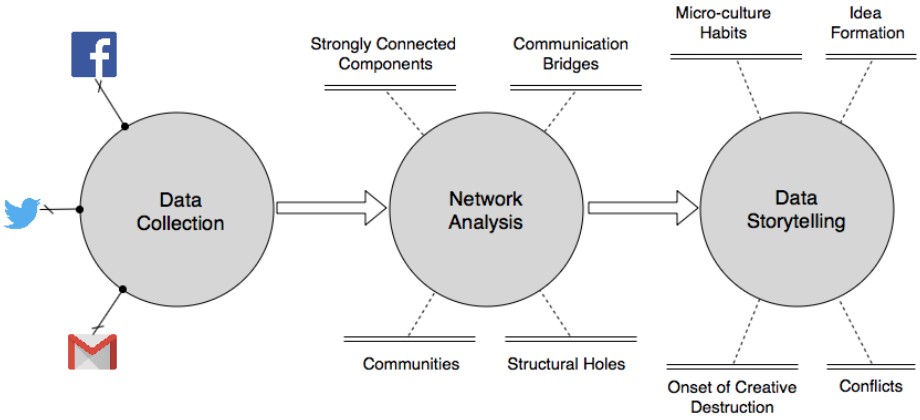

Fig. 1: End-to-end data analysis pipeline consisting of data collection, network analysis, and data storytelling.

2.2 System Overview

Figure 1 depicts the complete data pipeline. The pipeline consists of three main stages as data collection, network analysis, and data storytelling. Next, we discuss each stage in detail.

Data Collection Using Twitter API, we retrieve Twitter users of interest and their connections (the identities of followers and followees). After retrieval, we use our email directory, which contains user profiles for proper identification, in order to validate Twitter users in our community. For this purpose, we compute a text-based membership score based on Levenshtein distance, the value of which ranges from 0% to 100% [7]. All information retrieved from Twitter for each user along with this computed score is stored in a relational datastore. The

score is used in order to filter out all accounts with membership scores below a predetermined threshold τ. The default value of τ is 90%.

Algorithm 1 Crawler for identifying Twitter accounts from a given community.

Input: ego_nodes initialized with known accounts.

```
crawl_pool ← ego_nodes;
while crawl_pool is not empty do
    node ← pop node from the head of the pool;
    retrieve user_info, followers, followees of node from Twitter;
    for follower in node.followers do
        follower[membership] ← compute the membership score;
    end for
    node.followers ← {follower | follower[membership] ≥ τ};
    write node into the database;
    crawl_pool ← crawl_pool + node.followers;
end while
```

Similar to how ego networks are built, we begin our crawl starting from a select set of institutional Twitter accounts as ego nodes[1], which we know with utmost certainty that they belong to our university. These ego nodes are the main Twitter account, the Twitter accounts for all academic departments and graduate programs, and the Twitter accounts for all known student clubs. The followers of each ego node are added to the crawl pool for further exploration. The exploration is recursive with a cap on recursion depth. Once all nodes are explored in the crawl pool, the crawling terminates. The complete algorithm is given in Algorithm 1. For each node explored, its data profile including the connections, i.e., who follows the user (followers), and whom the user follows (followees) is written to the database for downstream analysis. The data exchange that occurs between the web services involved during a crawling session is depicted in Figure 2.

Each time we run the crawler, new community members as of the last run would be identified since the chances are higher for them to follow either one of the institutional accounts or at least follow an existing user who has already been explored and added into our database. The existence of a connection between a pair of nodes is a volatile entity: it may get added indicating the start of a relationship, and later on it may get deleted indicating the end of a relationship. And the number of state changes a relationship goes through could be arbitrary for any given pair of nodes. For simplicity and efficiency, we maintain the state of a connection as a binary piece of information, i.e., as being either present or absent. However, the evolution of the relationship can still be studied using the snapshots of the network, which are taken on different times.

[1] Node, account, and user all refer to the same entity in this context.

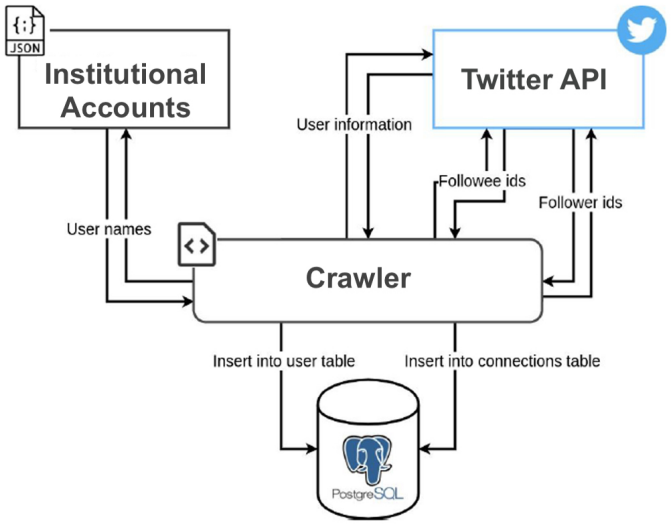

Fig. 2: The data collection from Twitter using its public API.

Network Analysis Each individual has a distinct set of personal characteristics. Each of us individually engage in a set of behaviours and activities that would drive the formation of links within the network, occasionally leading us to specific friendships that may defy any "norm". However in aggregate, links in a social network tend to form between people that are similar to one another [8]. That is, there are factors that exist outside the nodes and edges of a network (the network's structure), and that affect how the network's structure evolves. The compatibility of two individuals can strongly influence whether a link forms between them. This tendency is called homophily. Furthermore, shared activities between two individuals create an increased chance of an interaction between them, which would lead to the formation of a relationship between them in the future. Examples such of shared activities include working for the same company, living in a certain suburb, frequenting a particular pub, and playing golf in close proximity. These activities are "focal points of social interaction (foci)", i.e., social, psychological, legal, or physical mediums around which shared activities are organised [4]. In our setting, all of our institutional accounts constitute the set of foci. In order to assess whether homophily is prevalent in our university and how it manifests itself, we chose to study the natural language used in our community. The reason for this choice is two-fold: (i) we can determine the language of choice objectively and (ii) each language subject to our assessment is adopted by a sufficiently large number of users in the community.

Distinct closure processes drive the formation of individual links in a given network. There are three closure processes as shown in Figure 3. Membership closure (also known as social influence) is the tendency of a person to alter her behaviour in order to align it with those of her friends, who influence her

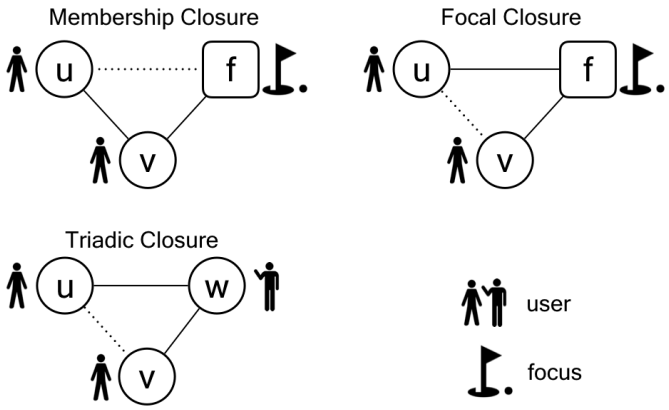

Fig. 3: There are three closure processes that operate on a network. In the drawings, the nodes u, v, and w represent distinct users while the node f represents a focus. Social influence drives membership closure; selection drives focal closure; trust and gravity drives triadic closure.

socially. Focal closure (also known as selection) is the tendency of a person seeking out to form friendships with people that are similar and compatible with herself. Triadic closure is the tendency of two people, who share a common friend, becoming friends due to the trust being established transitively and the facilitation of their friendship by the mutual friend.

The edges in a social-affiliation network are between two users, or between a user and a focus[2]. An edge $\{u, v\}$ is a focal edge if either u or v is a focus. The formation of a focal edge $\{u, f\}$ between a user u and a focus f at time t is because of the membership closure if there existed two edges before t, one of which is an edge between u and another user v, and a focal edge between user v and focus f. The formation of an edge $\{u, v\}$ between two users at time t is because of the focal closure if there existed a focus f with edges $\{u, f\}$ and $\{v, f\}$ being present before time t. The formation of an edge $\{u, v\}$ between two users u and v at time t is because of the triadic closure if there existed edges $\{u, w\}$ and $\{v, w\}$ before time t where w represents a user.

Data Storytelling We used Django as our web development framework for creating a dynamic and interactive dashboard [2]. A network, which is represented in javascript object notation (json) data format, defines the granularity of data exchanged between the network analysis layer and the data storytelling layer. Data scientists can use the interactive widgets provided in the dashboard in order to select nodes and edges of interest, which defines a new network to

[2] It is possible to have an edge between two foci in Twitter because the institutional accounts are maintained by human beings, who may also be "regular" users in the network.

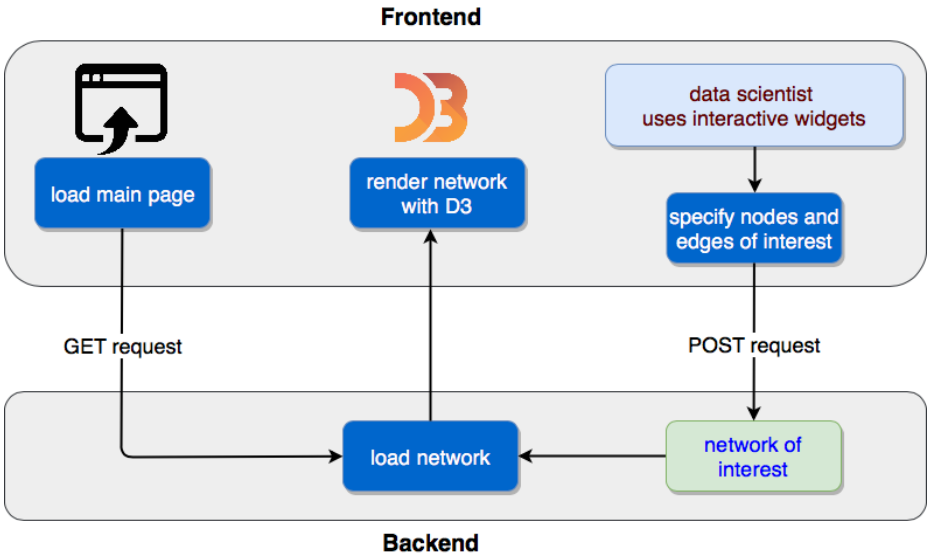

Fig. 4: The interface between the front-end and the back-end of our system.

render as shown in the right hand side of Figure 4. We used a popular javascript library called data driven documents known as *d3* for rendering the network within a browser as depicted in the left hand side of Figure 4 [1].

We used Girvan-Newman network partitioning algorithm in order to identify communities within the network and annotate user nodes with the communities they are associated with [5]. User nodes in the same community are displayed with the same color. When specifying the nodes and edges of interest, data scientists can use appropriate levers, which are added on during the structural analysis of the network. For example, one can filter nodes by their clustering coefficient in order to disclose tightly-knit regions, or cliques. The range of possible values can be used to specify the size of each node. For instance, a node with a high degree can be depicted larger than a node with a lower degree.

Each time the crawler is run, it takes a snapshot of the network. We run the crawler periodically so that we can study the evolution of the network. For this purpose, we added a scrollbar in the dashboard, which is used to access and render the state of the network over the snapshot timeline. Furthermore, the dashboard allows zooming in on a specific user node. Using a search query, data scientists can search for specific user/s of interest, and the rendering responds automatically. The matching set of user nodes to the query are highlighted with a higher opacity and a bigger size than the rest of the network. The network is centered automatically around one of the selected nodes (if there exists any).

3 Findings and Discussion

In this section, we present our initial findings on the social-affiliation network and the social network of our university.

3.1 Network Structure Analysis

We studied the structural properties of each network using a popular network analysis tool called NetworkX [9]. We computed graph modularity, graph diameter, and clustering coefficient per node. According to the snapshot taken on May 24^{th}, 2018, our networks have the "aggregate" properties (mean values) as shown in Table 1.

	Social-Affiliation Network	**Social Network**
Total number of nodes	1041	554
Total number of edges	2286	951
Node degree	4.39	3.43
Clustering coefficient	0.306	0.113
Modularity	0.4535	0.619
Graph diameter	7	8

Table 1: The mean structural properties of the social-affiliation network and the social network of our university computed using the data snapshot taken on May 24^{th}, 2018.

The social-affiliation network, which contains both users and their foci as nodes and the corresponding edges between these nodes, has a smaller diameter compared to the social network, which only contains users as nodes and the edges between different users. This is expected since the distance between two arbitrary users will be shorter when we include the possibly overlapping affiliations in the network as well. Similarly, the network with affiliations is more clustered having denser communities compared to the less clustered social network, which is devoid of any affiliations. The social-affiliation network is less modular with a higher average clustering coefficient and a smaller diameter compared to the social network, which confirms that it contains both dense and large communities.

3.2 Studying Homophily by Language

Between two snapshots taken on May 8^{th} and May 24^{th} respectively, we found that there was a total of 638 new focal closures. We define an edge $e = \{u, v\}$ between users u and v as heterogenous if their language of choice on Twitter is different. That is, u and v use different natural languages majorly to tweet. On the other hand, if u and v have adopted the same language, then an edge between them is considered as homogenous.

The degree of heterogeneity of a whole network is the ratio of heterogenous edges among all edges. We can compare the degree of heterogeneity "observed" to the theoretical level of heterogeneity "expected". This can be done by first computing the language distribution over the users in the actual network. Then, for any edge $e = \{u, v\}$, the languages adopted by its endpoints u and v are randomly assigned using the underlying language distribution. Assume that there are two possible languages L_1 and L_2 used in the community. Let the ratio of users that adopt L_1 be p, and the ratio of users that adopt L_2 be q. Then, homogenous edges in a theoretical network occur with a ratio of $p \times p = p^2$ and $q \times q = q^2$ for L_1 and L_2 respectively. An heterogenous edge $e = \{u, v\}$ in a theoretical network occur with a ratio of $p \times q + q \times p = 2pq$, where the first term on the left hand side of the equation represents u adopting L_1 and v adopting L_2, and the second term on the left hand side of the equation represents u adopting L_2 and v adopting L_1. The value of $2pq$ is the ratio of heterogenous edges expected in a network with an underlying language distribution parameterized by p and q. We compute the ratio r of heterogeneous edges in the actual network, and compare this empirical value to the theoretical limit $2pq$. If r is significantly less than $2pq$, then this observation constitutes an evidence for the presence of homophily in the community, for which the network stands for.

Table 2 summarizes the results we obtained on our community. As can be seen from the results, the ratio of heterogeneous edges in both networks we constructed are above their corresponding theoretical limits. This is an indication of a multi-cultural ambience being prevalent in our university. Note that the institutional accounts seem to have a more homogenous distribution weighing on one particular language, thereby reducing the ratio of heterogeneous edges in the social-affiliation network to 0.48 in comparison to the social network, which does not have any institutional Twitter accounts, and which has a higher degree of heterogeneity with the corresponding ratio being 0.52.

Network Studied	Ratio of Heterogeneous Edges	Theoretical Limit = $2pq$
Social-Affiliation Network	0.48	0.38
Social Network	0.52	0.39

Table 2: The study of homophily by language of choice in our networks. The observed ratios are compared with the theoretical limits. The results indicate that there is no homophily present in the cultural attribute we studied.

4 Conclusions

Working with the right context data, we captured valuable subjective characteristics of our university as a community. By taking into account the time dimension and the focal dimension, we were able to study the onset and the dynamics of implicit phenomena, such as the tendency of individuals to selectively

seek out new relationships. We studied homophily by the spoken and written language, and observed that there is no language barrier specifically before a true multi-cultural interaction and integration in our university.

In the future, we plan to enrich the data storytelling capability and capacity of our system. We have already started to work on how to adjust the level of resolution on the network with smart overlays, which hopefully could help us first detect certain calls for change in the community, and then monitor their progression. In order to synthesize the network in multi-resolution, the strongly connected components of the network will be identified; the directed acyclic graph will be overlaid on the components so as to better visualize the flow of influence and change. At the peripheries of micro communities, we plan to isolate the local bridges with a sufficiently large span, study the individual characteristics of their endpoints, and visualize what parts of the network they interconnect.

References

1. Bostock, M., Ogievetsky, V., Heer, J.: D3 data-driven documents. IEEE Transactions on Visualization and Computer Graphics **17**(12), 2301–2309 (Dec 2011)
2. Burch, C.: Django, a web framework using python: Tutorial presentation. Journal of Computing Sciences in Colleges **25**(5), 154–155 (May 2010)
3. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York, NY, USA (2010)
4. Feld, S.L.: The focused organization of social ties. American Journal of Sociology **86**(5), 1015–1035 (1981)
5. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99**(12), 7821–7826 (2002). https://doi.org/10.1073/pnas.122653799
6. Lu, Y., Wang, H., Landis, S., Maciejewski, R.: A visual analytics framework for identifying topic drivers in media events. IEEE Transactions on Visualization & Computer Graphics **24**(9), 2501–2515 (Sept 2018)
7. Miller, F.P., Vandome, A.F., McBrewster, J.: Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance. Alpha Press (2009)
8. Moody, J.: Race, school integration, and friendship segregation in america. American Journal of Sociology **107**(3), 679–716 (2001)
9. Schult, D.A.: Exploring network structure, dynamics, and function using NetworkX. In: In Proceedings of the 7th Python in Science Conference (SciPy. pp. 11–15 (2008)
10. Wu, Y., Chen, Z., Sun, G., Xie, X., Cao, N., Liu, S., Cui, W.: Streamexplorer: A multi-stage system for visually exploring events in social streams. IEEE Transactions on Visualization & Computer Graphics **24**(10), 2758–2772 (Oct 2018)