Mining Social Interactions in Privacy-preserving Temporal Networks

Federico Musciotto*, Saverio Delpriori†, Paolo Castagno‡ and Evangelos Pournaras§
*Dipartimento di Fisica e Chimica, Universitá degli Studi di Palermo, Palermo, Italy
federico.musciotto@unipa.it
†Universitá degli Studi di Urbino, Urbino, Italy
saverio.delpriori@uniurb.it
‡Universitá degli Studi di Torino, Torino, Italy
castagno@di.unito.it
§Professorship of Computational Social Science, ETH Zurich, Zurich, Switzerland
epournaras@ethz.ch

Abstract—**The opportunities to empirically study temporal networks nowadays are immense thanks to Internet of Things technologies along with ubiquitous and pervasive computing that allow a real-time fine-grained collection of social network data. This empowers data analytics and data scientists to reason about complex temporal phenomena, such as disease spread, residential energy consumption, political conflicts etc., using systematic methologies from complex networks and graph spectra analysis. However, a misuse of these methods may result in privacy-intrusive and discriminatory actions that may threaten citizens' autonomy and put their life under surveillance. This paper studies highly sparse temporal networks that model social interactions such as the physical proximity of participants in conferences. When citizens can self-determine the anonymized proximity data they wish to share via privacy-preserving platforms, temporal networks may turn out to be highly sparse and have low quality. This paper shows that even in this challenging scenario of privacy-by-design, significant information can be mined from temporal networks such as the correlation of events happening during a conference or stable groups interacting over time. The findings of this paper contribute to the introduction of privacy-preserving data analytics in temporal networks and their applications.**

I. Introduction

The introduction of Internet of Things technologies along with advances in ubiquitous and pervasive computing has brought paramount opportunities for data collection and analysis. This is especially the case for social interactions in domains such as healthcare [1], conflicts [2], disease spread [3] and other [4]. Mobile phones, environmental sensors, and social networks can track human mobility, relationships and their significance, especially when the interaction dynamics are represented with temporal networks that model the evolution of interactions [5]. Such networks can encode at a fine-grained granularity personal information sensitive to citizens. Mining this sensitive information raises serious privacy threats and opportunities for discriminatory and surveillance actions [6] that have significant implications on the autonomy of citizens [7]. However, if the sparsity of the data increases as a result of self-determined decisions that improve the privacy of citizens, the effectiveness of mining social interactions over temporal networks come in questions. This paper addresses on this challenge. It applies techniques that can measure the correlation of events and can detect stable groups evolving over time within a temporal network of social interactions.

This paper focuses on privacy-preserving interactions derived by the physical proximity of agents in a public space such as a conference event. When agents are in close proximity at a specific time point, an interaction can be defined and modeled by an edge interconnecting the two agents, the nodes of the social network. The edges interconnecting the agents at each time point/window form together the temporal network studied. Privacy-intrusion comes in question when the localization of the agents in the public space and the calculation of proximity are absolute such as the case in which nodes and edges are computed based on geolocated data, e.g. GPS. In contrast, when (i) the proximity data shared are anonymized, (ii) data sharing is self-determined by the agents and (iii) the localization of interactions relies on relative distances using technologies such as bluetooth beacons, privacy-preservation increases [8] at a cost of significantly lower data quality that challenges the analysis of temporal networks. This paper studies a real-world deployment of such a privacy-preserving social network formed by a highly-concerned community about privacy: the 2014 Chaos Communication Congress in Hamburg. Data are collected with the Nervousnet platform [9] from 154 users during the 4 days of the congress. The technology used includes bluetooth beacons carried and distributed in the physical space by the participants themselves. It also involves a mobile app that participants had to download and use during the conference and a web server to collect the data. The experimental analysis of the temporal network shows high correlations of the graphs within each scheduled event during the congress, however these correlations decrease when comparing graphs among different scheduled events. This suggests that every event has different interaction patterns and dynamics. Moreover, the proposed analysis detects small, yet statistically significant, stable groups that emerge within the temporal network studied.

The rest of this paper is organized as follows: Section II introduces the computational model of sparse temporal networks. Section III illustrates the data acquisition and analysis process and the results on events correlation and detection of stable social groups. Section IV reviews related work. Finally, Section V concludes this paper and outlines future work.

II. A Computational Model for Sparse Temporal Networks

This section proposes a model designed for representing a social system in which the information related to the interactions among agents is sparse. Such a system can be represented by an undirected graph.

A graph G is a mathematical object defined by the ordered pair (V, E), where V is the set of nodes and E the set of edges. In graphs, which describe social systems, the nodes may represent the agents, while the edges represent the interactions among them. More detailed information about graphs can be found in [10]. Temporal networks, instead, are graph in which the interactions among agents are not fixed and may vary at different time points, for example as a consequence of the agents' mobility. This temporal structure can affect the dynamics of phenomena, such as the diffusion of information, since the topological properties of the graph change. More information on temporal graphs can be found in [5].

In this paper, interactions are defined by the physical proximity of agents. Measures of proximity can be obtained by detecting the presence of other agents in their neighborhood. Internet of Things technologies, such as bluetooth beacons and pervasive/ubiquitous sensors can track proximity in real-time by measuring the signals strength. Proximity data are collected by several applications, however proximity data remain sensitive information. Self-determination of sharing user data is a requirement in several of these applications for meeting privacy requirements and therefore, the sparsity of the collected proximity data challenges data analytics. GPS sensor data is an example as often agents are not willing to be tracked at fine-grained granularity.

The objective of the proposed model is the analysis of social event and group dynamics in social systems governed by interactions defined by the proximity of the participating agents. This analysis involves the following two aspects: (i) changes in the volume of interactions, estimated on the basis of correlation coefficients and similarity indices computed on temporal graphs extracted from proximity data. (ii) the detection of agent groups with stable observed interactions.

The adoption of privacy-preserving methods [11] that allow agents to determine when and to what extent they share proximity data requires an analysis of the quality of the resulting datasets, in order to quantify their degree of sparsity. Intuitively, such a quality measure should quantify to what extent the agents are on average active in the system: the sparser is the dataset, the fewer agents are recorded as active in several different time points. To this extent the following approach is introduced: first, the total time period spanned by the system is splitted into non-overlapping time windows of equal size. Then, the quality index

$$q(N) = \frac{1}{N}\sum_{i=1}^{N} A_i/P \tag{1}$$

is computed, with N the number of time windows, A_i the number of agents in the time window i and P the number of agents tracked at least once in the whole studied dataset. The quality index $q(N)$ depends on the size of the N time windows. This observation suggests checking the dataset for different values of N in order to find the time scale with the highest quality along with the desired time resolution. Indeed, looking at how the quality index varies at different values of N gives a general hint on the overall sparsity of the data, but implies a clear trade-off: the longer the time window in which the quality index is computed, the higher amount of information is aggregated per time window and the lower the time resolution is.

A. Temporal network events

When analyzing a temporal network that models a social event with significant changes during scheduled moments, it is often convenient to treat time as a non homogeneous variable [12]. For example, it is expected that only a few, or even no agents interact with others at night during a congress, or that the interaction dynamics during a scheduled pause in the social activities is quite different from what observed when they are in progress.

For this reason, a temporal partition of the system that does not consider time as a homogeneous variable is considered here. From now on, an event $H(T, t)$ is defined as the set of graphs extracted by three time windows of width t centered in $T - t$, T and $T + t$, where T is the timestamp of one of the scheduled activities of the social phenomenon under analysis. The three graphs in each windows are referred to as the *earlier*, the *current* and the *following* graph respectively. Figure 1 illustrates the time resolution of an event.

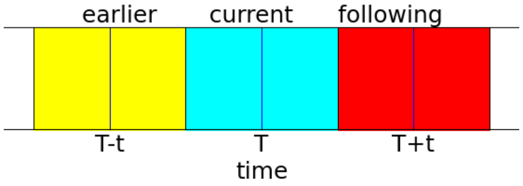

Figure 1: Visualization of the time span defined for events. The time interval filled in yellow represents the *earlier*, the cyan one represents the *current* and the red one represents the *following* graph.

Each graph is obtained by aggregating all the interactions occurring among the agents in each window. If the proximity data under analysis produce bipartite graphs, i.e. graphs in which there are two clear distinguished sets of agents, with no interactions among the agents of the same set, it is possible to extract the projection graph on one of the two sets. The projection is performed by adding an interaction among each pair of agents of the chosen set if and only if they are connected to at least one common agent of the other set in the bipartite network [10]. Bipartite networks can occur when agents do not interact directly with each other, but only with a set of static sensors. In this case of the projected network, two agents are connected by an interaction if they are both interacting with the same sensor.

For both bipartite or monopartite graphs, the multiple interaction edges among the same pairs of agents are removed, in order to obtain a simple graph. In this way, the heterogeneity and the noise in the data are partially reduced.

B. Events analysis and dynamics

Once an event is defined and extracted, the behavior of the agents tracked within each time window can be analyzed. One interesting aspect of this behavior is related to the dynamics of agent interactions: measuring whether their volume and their flux are stable or not within one or different events provides a useful insight on the attributes of the events. As an example, one could expect that the crowds attending music concerts of the same genre have similar patterns of interactions, and vice versa. This section introduces a set of tools to analyze these patterns.

Since an event is defined as a collection of networks, its analysis exploits the tools of graph theory. In order to characterize the activity of the agents, the degree sequence is extracted from each graph. Degree centrality is a radial centrality measure [13]: it takes into account the walks or paths which start/end in the selected nodes, in contrast with the medial class of centrality measures, which instead considers the walks or paths which pass through the nodes. Thus, degree centrality seems the most suitable choice to track the number of interactions observed among the agents. Another relevant alternative is the spectral radius of the adjacency matrix. Starting from the degree sequence of the graphs that characterize the events, a vector whose size is equal to the number of all the agents tracked at least once in the three subwindows is assigned to each graph of the event. Each entry has the value of the degree of the corresponding agent if it is active in the selected time window, 0 otherwise. Then, the similarity between each pair of vectors is evaluated. Two similarity and correlation measures are proposed here due to their large applicability: the generalized Jaccard index [14] and the Spearman's rank coefficient [15].[1].

Moreover, from the Jaccard index J, it is possible to obtain a distance d_J between any graphs pair (i, j) [16], $d_J(i, j) = 1 - J(i, j)$. The distance d_J is a metric that fulfills the triangle inequality. Thus, computing the differences $d_J(e, c)+d_J(c, f)-d_J(e, f)$, where e, c, f stand for *earlier*, *current* and *following*, is a way to evaluate whether the low quality of data is not causing inconsistencies: if the triangle inequality $d_J(e, f) \leq d_J(e, c) + d_J(c, f)$ is not fulfilled, the sparsity of data alters the expected distribution of interactions in time by making the presence of agents discontinuous.

A different analysis starting from the proposed formalism can be performed. It concerns the similarity of event pairs, starting from the triplet of degree sequences defined above. The similarity measure between events A and B is defined as follows: for both events the monopartite or the projected network is considered and the triplet of vectors reporting the degree sequence is extracted as shown above. The averages, $\tilde{D}_A$ and $\tilde{D}_B$, and the coefficients of variation [17], CV_A and CV_B are computed: the first ones are indicators of the amount of interactions per agent, while the second ones provide information about the changes occurring in the same quantity. The similarity between the couples ($\tilde{D}_A$, $\tilde{D}_B$) and (CV_A, CV_B) is computed using the generalized Jaccard index and the Spearman's rank coefficient. In this way, global correlation matrices that quantify the similarity among all the considered events are obtained. These matrices provide a general overview on the overall system dynamics, showing whether the flux of interaction per agent undergoes significant changes during the evolution of the social system.

C. Stable group detection

A significant attribute of complex networks is their community structure [18]. A community is a subset of strongly interconnected nodes with a higher number of interactions among its components than with external nodes. Thus, community detection can show how a global set of agents is partitioned in clusters related to different tasks within a complex system. Historically, community detection was born within the field of social studies [19], but it expands to biology [20], informatics [21], virtual social mining [22] and other fields [18]. Many different contributions are shaping the community detection field towards different directions [23] [24]. In the case of a social system represented by proximity data, detecting communities involves looking for clusters of agents that develop strong, close interactions among themselves. The presence of low quality data, though, makes the detection of a valid community structure a challenging task due to the presence of missing interactions, coarse-grained resolution and other sources of sparsity. In this case, the detection of small, statistically significant groups that do not cover the whole population is a more relevant semantic in this context. For this reason, the rest of this paper refers to groups rather than communities, though the techniques discussed originate from techniques on community detection.

Group detection begins with the aggregated graph of all occurred interactions, or its projection on one of the two sets, if the graph is bipartite. In light of what already discussed about low quality data, the aggregated graph may not be fully representative of the system. There may be many agents tracked for a limited time. Another issue is the choice of not being tracked that each agent can make at any time when using a privacy-preserving platform. This introduces a significant degree of heterogeneity that can misrepresent the detected groups. In order to deal with these challenges, a suitable time scale and a filtration procedure for interactions is introduced. The intuition is that, once all interactions are aggregated, stable groups can be detected only after filtering out the statistically less significant information.

To this extent, the proposed method filters out all "weak" interactions among agents. The usage of a sliding time window within which all interactions among the same two agents are treated as a single one is proposed as a measure of "strength". Agents with multiple interactions in different time windows are described as two nodes connected by an undirected weighted edge whose weight corresponds to the number of time windows in which at least one interaction is registered. In this way the multiplicity of interactions per agent is reduced.

At this point, the network has to be filtered. The definition of interactions strength depends on the length of the time windows. Moreover, in order to filter out the weaker interactions, a threshold on the strength is required. Thus, the length of the time windows and the strength threshold are two parameters on which the topology of the filtered graph depends. In order to detect the best group partition, a large number of parameter couples is generated. For each parameter couple, the graph is filtered and a community detection algorithm is applied on it. All the resulting partitions are ranked by using modularity, a commonly used measure for graph partitions [18]. Finally, the best performing partition is selected.

A robust statistical validation is required for the evaluation of the detected groups. When dealing with complex systems represented as graphs, a well-known validation process is the usage of random Erdos Renyi graphs as null models. The properties of Erdos Renyi graphs can be computed analytically in most cases [10]. Moreover, in order to obtain more statistics on real data, an equivalent family of graphs can be generated in order to perform the test. Such an equivalent class of graphs is obtained by rewiring the interactions among agents without changing their degree sequence, according to the configuration model [10], [25].

III. Data Analysis and Evaluation

Analysis and evaluation is performed over a proximity dataset collected during the Chaos Communication Congress[2] in Hamburg, Germany, in 2014. 120 plastic cases encapsulating bluetooth beacons were 3D printed at the Professorship of Computational Social Science at ETH Zurich in Switzerland. 20 of these beacons were static and placed at specific rooms of the congress to localize room-specific interactions. The rest of the beacons were distributed to the congress participants. Bluetooth signals were captured by the Nervousnet platform implemented as a mobile app [9]. 154 unique participants downloaded the app during the congress. Thus, the app records bipartite networks in which the projection is performed on the set of participants who downloaded it, in order to consider only human agents. The app is capable of measuring the Received Signal Strength Indication (RSSI) and TX Power of the signal emitted by the bluetooth beacons. Based on this information the physical distance

[1] It must be noticed that this methodology is able to track the changes only in the number of interactions per agents and does not consider if the counterparts of the interactions have changed as well.

[2] Avaliable at https://events.ccc.de/congress/2014/Fahrplan/schedule.html (last accessed: June 2016)

ID event	Day	Time	Type of transition
1	2014-12-27	15:00	Conference/Break
2	2014-12-27	16:00	Break/Conference
3	2014-12-27	19:30	Conference/Break
4	2014-12-27	20:30	Break/Conference
5	2014-12-28	15:00	Conference/Break
6	2014-12-28	16:00	Break/Conference
7	2014-12-28	19:30	Conference/Break
8	2014-12-28	20:30	Break/Conference
9	2014-12-29	15:00	Conference/Break
10	2014-12-29	16:00	Break/Conference
11	2014-12-29	19:30	Conference/Break
12	2014-12-29	20:30	Break/Conference
13	2014-12-30	15:00	Conference/Break
14	2014-12-30	16:00	Break/Conference

Table I: Time schedule of events during the 2014 Chaos Communication Congress.

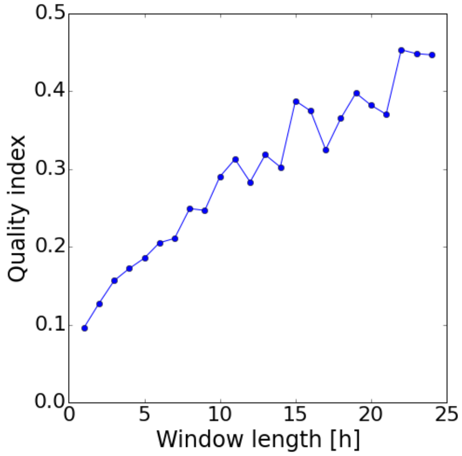

Figure 2: Quality index as a function of the length of the time windows.

r between a participant with the bluetooth beacons transmitter and a participant with the Nervousnet mobile app can be derived as follows:

$$r = \sqrt{10^{(TX-RSSI)}/10}, \tag{2}$$

where r is expressed in meters [26]. Edges are counted when distances are lower than 5 meters suggesting an actual social interaction. There was no GPS information recorded at any time of the congress. The data collected are anonymized and timestamped. Congress activities concern lectures and workshops, with a maximum of four different activities running in parallel. This analysis focuses on transitions from a conference event to a break, and vice versa, scheduled throughout the congress. The timestamps of the transitions are illustrated in Table I.

The quality index of the dataset is illustrated in Figure 2, in which the size of the time windows varies from 1 to 24 hours, while the quality index is always smaller than 0.5. This shows that, on average, 50% of all the agents are not continuously present throughout time.

Figure 3, instead, shows the percentages of agents observed over time for different number of recorded interactions during the events extracted from Table I. Figure 4 shows the same percentages heatmap computed on time windows which equally span the duration of the whole congress. It is evident that in the first figure the heterogeneity is reduced and the distributions of percentages are more homogeneous.

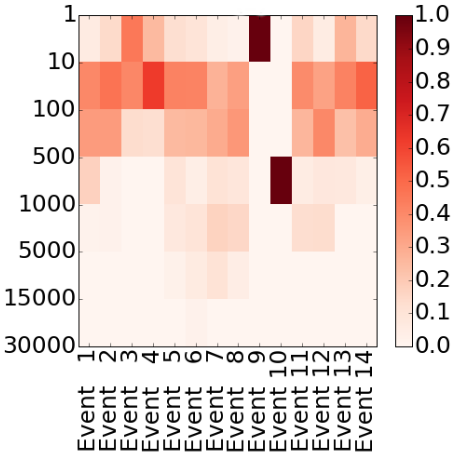

Figure 3: Heatmap with the percentages of agents performing a number of interactions contained in one of the intervals spanned on y axis during the different events listed in Table I. All columns sum to 1.

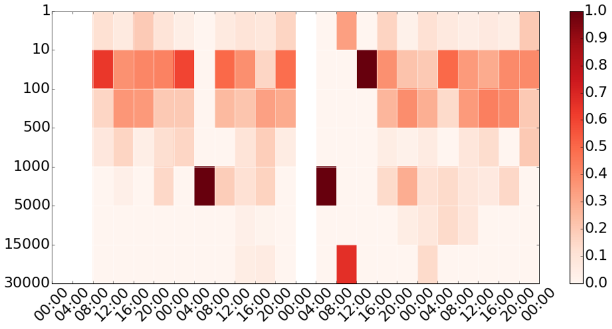

Figure 4: Percentages heatmap built on windows spanning equally to the whole duration of the congress. White columns represent time windows in which no interactions were recorded.

Figure 5 shows in blue the decumulative distribution function (ddf) of the degree sequences aggregated on all bipartite networks. In the same figure, the red line represents the ddf of the degree sequences on the projected networks of the same events. The second distribution is much slimmer than the first one: this means that in the projected networks agents have lower degrees.

A. *Events correlation*

Figures 6 illustrates the similarity measures between the degree sequences of the *earlier*, the *current* and the *following* graphs (e,c,f) for all non-empty events $\{H(T,t)\}$. Table I indicates the values of T used and $t = 0.5$ hours, which allows to consider the maximum possible size of non-overlapping time windows. It should be noticed that, apart from few exceptions, the values of similarity are high when using the Jaccard index: almost 60% are above 0.5. A high value of similarity indicates that the system does not go through significant changes, i.e. a large number of agents have a comparable amount of interactions during the event. The agent interactions remain stable within the same event. On the other hand, since the timestamps considered in Table I are transition points from different activities to scheduled pauses, the low correlation values suggest that agent interactions change.

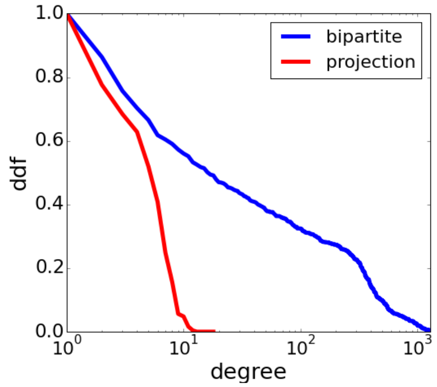

Figure 5: The decumulative distribution function of the degree sequences aggregated on all events represented by bipartite graphs (in blue) and by projected graphs (in red).

Moreover, the triangle inequalities of the Jaccard index on the triangles given by the vertices (e,c,f) are computed for each event of Table I by evaluating the differences $d_J(e,c)+d_J(c,f)-d_J(e,f)$. The results are shown in the yellow bars of Figure 6a. The high values of the bars show that the proposed partition of time is well-defined: in the temporal segments $\overline{ec}$, $\overline{cf}$ and $\overline{ef}$ there are no inconsistencies caused by the missing data in the dataset due to the privacy-preserving design of Nervousnet.

The outcome of similarity between events pairs are illustrated in Figures 7 as correlation matrices. In this case, the temporal networks have low values of similarity: almost 60% of all event pairs have a correlation lower than 0.2. Moreover, the higher similarity values, from 0.2 to 0.6, are detected among pairs of events that happened at the same day. The choice of degree averages or degree coefficient of variations does not affect significantly the outcomes, whereas using the Jaccard index provides lower values of similarity.

The outcomes of this measure are a complement of what already shown in Figures 3-4. Indeed, besides confirming the heterogeneity and sparsity of the data, this similarity measure reveals that the volume of interactions per agent during the different events shows significant changes. This underlines that the most relevant factor, when characterizing the time evolution of participation in the congress is time rather than the event topics: the same agents are more likely to be tracked in events that are close in time rather than in events that have similar topics.

B. Detection of stable groups

Figure 8 shows the number of occurrences for each agent in different events. It is evident that only a low fraction of agents are active in more than two events. This justifies the idea of detecting stable clusters in the whole dataset after applying a filtration procedure of the system within a suitable time scale. In this way, the most statistically relevant information is extracted from the system for the group detection.

In order to find the most proper agent grouping of the system, one or more community detection algorithms are applied on the graphs obtained through 10000 couples of randomly generated parameters. Two of the most promising community detection algorithms [27] are employed in this context: the first one is the multilevel modularity optimization algorithm [28] and the second one is the InfoMap algorithm [29]. Each parameters couple is composed of the length of the time window, ranging from 1 second to 1 day, and the threshold of strength used to filter the graph interactions, ranging from 1 to 20: each interaction, whose strength is lower than the threshold is filtered out. The groups with the highest modularity are detected by applying the multilevel algorithm on a network obtained with time windows of about 8.5 hours (30180 seconds) and with a threshold of 10 for the weight of interactions. The results of this procedure are shown in Figure 9. Thus, on the basis of the features in the resulting graph, it can be confirmed that only few agents have stable and continuous interactions: the filtering procedure is severe and leaves only 9 nodes, i.e. 8% of the total, which are partitioned in three groups.

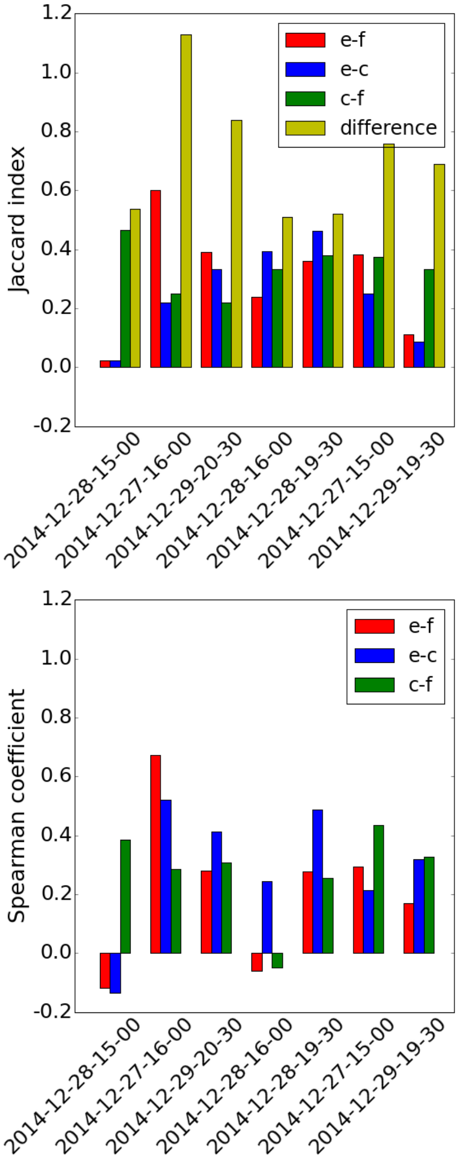

Figure 6: Bar chart representation of (a) the generalized Jaccard index and (b) the Spearman's rank coefficient between the degree sequences of the *earlier*, the *current* and the *after* graphs (e,c,a) of all non empty events. In (a), yellow bars illustrate the differences $d_J(e,c)+d_J(c,a)-d_J(e,a)$.

In order to evaluate the relevance of the structural properties emerged from the group detection procedure, the same procedure has to be applied to a class of random Erdos Renyi graphs [30].

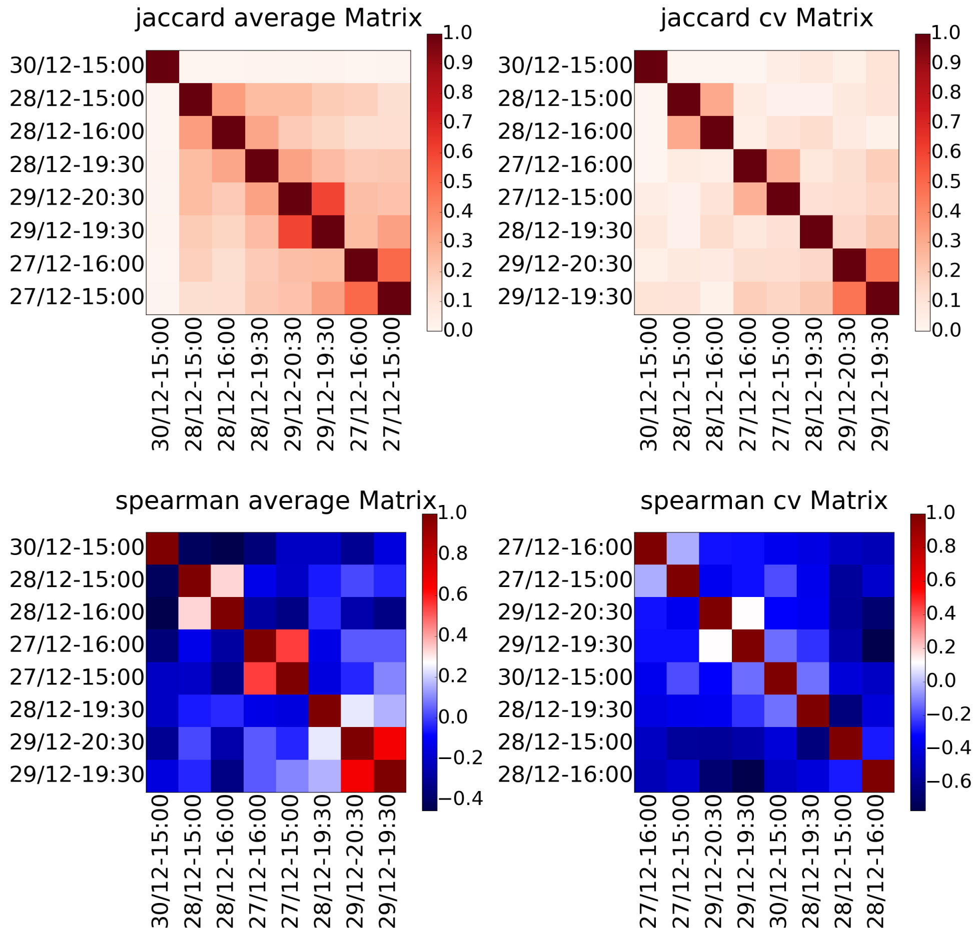

Figure 7: Correlation matrices computed using Jaccard's and Spearman's coefficients on the averages and coefficients of variation for the degree sequences of events pairs.

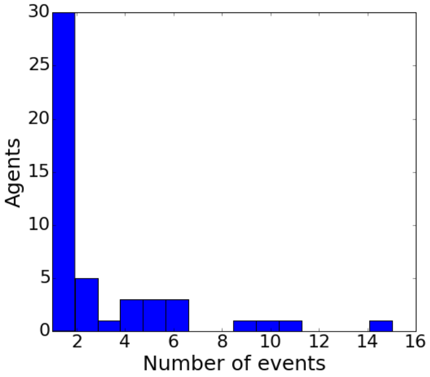

Figure 8: Number of agents in different events.

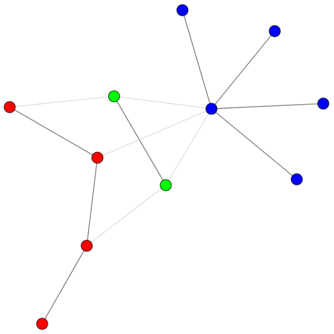

Figure 9: Result of the group detection on the filtered graph: the different colors represent the groups detected after the filtration procedure

This class is used as a null model, while a class of equivalent graphs obtained according to the configuration model [31], which preserve the degree distribution of the original graph, is used as the class to be tested. It is easy to verify that the number of groups in the set of configuration model graphs is almost the same: variance and standard deviation confidence intervals are tight, as shown in Table II.

Comparing the results on the two different sets, the set of random graphs is characterized by a higher average and a higher dispersion of the number of groups, as reported in Table III. This shows that the final results on the two sets diverge: the Erdos Renyi graphs, taken as a null model, have a different grouping than the graph extracted from the empirical data. This is an indicator of the statistical significance of the model outcome.

	Lower	Center	Upper
Average	3.19	3.19	3.19
Variance	0.16	0.16	0.16
Standard Deviation	0.40	0.40	0.41

Table II: Results of group detection on the configuration model graphs with a confidence interval $\alpha = 0.99$. Average, variance and standard deviation are computed on the number of groups detected in different graphs.

	Lower	Center	Upper
Average	3.53	3.54	3.55
Variance	0.43	0.44	0.45
Standard Deviation	0.66	0.66	0.67

Table III: Results of group detection on the Erdos Renyi graphs with a confidence interval $\alpha = 0.99$. Average, variance and standard deviation are computed on the number of groups detected in different graphs.

IV. Review of Related Work

Physical proximity in closed and open air venues can be measured via wireless beacons, wearables and mobile devices for detecting social interactions among people equipped with these devices [33]–[36]. In particular, the topic of conference proximity analysis is studied by Isella et al. [37] through a detailed analysis of face-to-face contact networks under two different scenarios: a scientific conference and a long-running museum exhibition. The former case concerns a 'closed' system in which a group of individuals interacts in a repeated pattern. The latter case concerns an 'open' environment with a flow of individuals streaming through a baseline. Face-to-face proximity data are collected by means of RFID badges carried by attendees. An extensive analysis on static and dynamic properties of the proximity network is performed. In contrast, this work relies on proximity data collected via bluetooth beacons that allow measurements of longer distances with cheaper equipment. The proposed analysis does not require any absolute localization of the interactions, all computed distances between agents are relative to each other.

Similarly, Cattuto et al. [38] analyzed the network dynamics of person-to-person interaction networks by collecting data from an office environment and an academic congress. They conclude that the node strength, represented by the sum of contacts duration of those nodes, grows super-linearly with the degree. Moreover, Barrat et al. [39] utilize a face-to-face contact network during a conference in conjunction with data collected by the Live Social Semantics framework [40], in order to analyze contacts patterns and the impact of parameters such as seniority and role. They compared virtual interaction networks, including friendships on online social networks, co-authorship networks, and face-to-face contact networks during a conference and they conclude to a clear emerging behavior according to which people tend to mix with other ones of similar scientific seniority levels. In contrast to these approaches, this paper focuses on studying complex dynamic interactions such as conference events, via temporal networks. New measurements and insights, such as the similarity and correlation of interactions within and across events, can be determined when social interactions are modeled via temporal networks.

As the tools of social network analysis advance, the amount of information that can be accurately inferred from raw and noisy data increases [41]. Privacy-preserving methods are studied and adopted to protect sensitive information about users and their social relationships. The state-of-the-art anonymization methods are categorized into three main categories [11]: (i) k-anonymity based privacy-preservation via edge modification, (ii) probabilistic privacy preservation via edge randomization and (iii) privacy preservation via generalization. Zou et al. [42] introduces a technique based on k-automorphism that permits network analysis over anonymized datasets. In contrast, Nervousnet combines anonymous record data with real-time self-determination of information sharing and therefore the privacy-preservation, from the perspective of each individual, is enhanced.

A different approach is adopted in [32]. In this work, the authors investigate a database containing GPS data on a set of vehicles. They observe that it is quite straightforward to extract from raw data sensitive information on the identity and the routines of several drivers. To address this issue, they propose a transformation of the original data through a Voronoi tessellation of the whole space. They show that their method makes the decrypting probability drop significantly, without destroying relevant collective properties of the system. In contrast to this paper, their limitation here is that some interesting data features such as the structure of communities made by drivers with similar profiles can be lost after such a transformation.

Cormode et al. [43] study networks as bipartite-graphs in which links are considered sensitive information that needs to be protected. They provide an anonymity method that hides this information while also preserving the graph structure. Another technique is introduced by Das et al. [44], who consider the problem of anonymizing the weights of the edges in a social network. They define a linear property that can be expressed by a specific set of linear inequalities of the edge weights, therefore they propose a framework to re-assign weights to edges by preserving a certain linear property. Empirical evaluation shows that this approach improves the edge k-anonymity of the modified graph and prevents the identification of an edge by its weight. In contrast to these anonymization approaches, the proposed techniques of this paper are applied to shared data exclusively self-determined by individuals via privacy-preserving platforms such as the Nervousnet. In other words, the data shared are collected by design in a privacy-preserving way.

V. Conclusion and Future Work

This paper concludes that mining of social proximity modeled via privacy-preserving temporal networks is feasible. Although temporal social networks encode additional complexity and require advanced techniques for analysis, this complexity turns out to enhance the data mining of social phenomena even when data are sparse due to privacy-preservation. This paper illustrates an analysis of a temporal network formed by privacy-preserving proximity data collected via the Nervousnet platform during the 2014 Chaos Communication Congress. Although the data are highly sparse, the proposed analysis shows high correlation values between network snapshots within an event, in contrast to low correlation values between network snapshots among different events. Moreover, groups with stable interactions and high modularity scores could be reliably detected from the highly-sparse dataset, however, the filtration process is severe and indicates small groups.

Future work includes the applicability of the proposed computational model to other such highly-sparse datasets but larger ones that allow the analysis of more complex social phenomena. Other methods for predictions and deep learning of privacy-preserving social interactions such as recurrent neural networks can be evaluated for this purpose.

Acknowledgment

The authors would like to alphabetically thank Patrick Misteli, Sam Sulaimanov, Siddhartha and the rest of the Nervousnet development team for their work on the platform development and data acquisition at the 2014 Chaos Communication Congress in Hamburg, Germany. This work is supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem' (http://www.sobigdata.eu) and the European Community's H2020 Program under the scheme 'ICT-10-2015 RIA', grant agreement #688364 'ASSET: Instant Gratification for Collective Awareness and Sustainable Consumerism'.

References

[1] Duncan Chambers, Paul Wilson, Carl Thompson, and Melissa Harden. Social network analysis in healthcare settings: A systematic scoping review. *PLoS ONE*, 7(8), 8 2012.

[2] Afndreas Wimmer and Brian Min. From empire to nation-state: Explaining wars in the modern world, 1816-2001. *American Sociological Review*, 71(6):867–897, 12 2006.

[3] Mark DF Shirley and Steve P Rushton. The impacts of network topology on disease spread. *Ecological Complexity*, 2(3):287–299, 2005.

[4] Xiangrong Wang, Evangelos Pournaras, Robert E Kooij, and Piet Van Mieghem. Improving robustness of complex networks via the effective graph resistance. *The European Physical Journal B*, 87(9):1–12, 2014.

[5] Petter Holme and Jari Saramäkid. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.

[6] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery*, 29(6):1733–1782, 2015.

[7] Dirk Helbing and Evangelos Pournaras. Society: Build digital democracy. *Nature*, 527(7576):33–34, 2015.

[8] Evangelos Pournaras, Jovan Nikolic, Pablo Velásquez, Marcello Trovati, Nik Bessis, and Dirk Helbing. Self-regulatory information sharing in participatory social sensing. *EPJ Data Science*, 5(1):1, 2016.

[9] E. Pournaras, I. Moise, and D. Helbing. Privacy-preserving ubiquitous social mining via modular and compositional virtual sensors. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, pages 332–338, March 2015.

[10] M. Newman. *Networks: An Introduction*. OUP Oxford, 2010.

[11] Xintao Wu, Ying Xiaowei, Liu Kun, and Chen Lei. *Managing and Mining Graph Data*, chapter A Survey of Privacy-Preservation of Graphs and Social Networks, pages 421–453. Springer US, 2010.

[12] Pitirim A. Sorokin and Robert K. Merton. Social time: A methodological and functional analysis. *American Journal of Sociology*, 42(5):615–629, 1937.

[13] Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, 2006.

[14] Jr. Sepkoski, J.John. Quantified coefficients of association and measurement of similarity. *Journal of the International Association for Mathematical Geology*, 6(2):135–152, 1974.

[15] J. L. Myers and A. D. Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, New Jersey, 2003.

[16] M. Levandowsky and D. Winter. Distance between sets. *Nature*, 234:34–35, 1971.

[17] B S Everitt and A Skrondal. *The Cambridge Dictionary of Statistics; 4th ed.* Cambridge University Press, Leiden, 2010.

[18] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[19] R. Boudon and J. S. Coleman. Introduction to mathematical sociology. *Revue française de sociologie*, 7(1):98–101, 1966.

[20] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, September 2006.

[21] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–71, March 2002.

[22] Rémy Cazabet, Hideaki Takeda, Masahiro Hamasaki, and Fréderic Amblard. Using dynamic community detection to identify trends in user-generated content. *Social Network Analysis and Mining*, 2(4):361–371, juin 2012.

[23] MEJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):36104, 2006.

[24] Rossetti G, Pappalardo L, and Rinzivillo S. A novel approach to evaluate community detection algorithms on ground truth. In *7th Workshop on Complex Networks*, Dijon, France, 2016. Springer-Verlag, Springer-Verlag.

[25] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. In *ADVANCES IN PHYSICS*, 2005.

[26] S. Virtanen. *Adoption and Optimization of Embedded and Real-Time Communication Systems*. Information Science Reference, 2013.

[27] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[28] Vincent D Blondel, Jean loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks, 2008.

[29] M. Rosval and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, 105:1118–1123, 2008.

[30] P. Erdös and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[31] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *ANNALS OF COMBINATORICS*, pages 125–145, 2002.

[32] Anna Monreale, Salvatore Rinzivillo, Francesca Pratesi, and Dino Giannotti, Foscaand Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3(1):1–26, 2014.

[33] Emiliano Miluzzo, Nicholas D. Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B. Eisenman, Xiao Zheng, and Andrew T. Campbell. Sensing meets mobile social networks: The design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, SenSys '08, pages 337–350, New York, NY, USA, 2008. ACM.

[34] Trinh Minh Do and Daniel Gatica-Perez. Human interaction discovery in smartphone proximity networks. *Personal Ubiquitous Comput.*, 17(3):413–431, March 2013.

[35] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.

[36] Niklas Palaghias, Seyed Amir Hoseinitabatabaei, Michele Nati, Alexander Gluhak, and Klaus Moessner. A survey on mobile social signal processing. *ACM Comput. Surv.*, 48(4):57:1–57:52, March 2016.

[37] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *CoRR*, abs/1006.1260, 2010. informal publication.

[38] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5(7):e11596, July 2010.

[39] Alain Barrat, Ciro Cattuto, Martin Szomszor, Wouter Van den Broeck, and Harith Alani. Social dynamics in conferences: analyses of data from the live social semantics application. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.

[40] Harith Alani, Martin Szomszor, Ciro Cattuto, Wouter Van den Broeck, Gianluca Correndo, and Alain Barrat. *The Semantic Web - ISWC 2009: 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, chapter Live Social Semantics, pages 698–714. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[41] Sugihara G, May R, Ye H, Hsieh C, Deyle E, Fogarty M, and Munch S. Detecting causality in complex ecosystems. *Science*, 496(338), 2012.

[42] Lei Zou, Lei Chen, and M. Tamer Özsu. K-automorphism: A general framework for privacy preserving network publication. *Proc. VLDB Endow.*, 2(1):946–957, August 2009.

[43] Graham Cormode, Divesh Srivastava, Ting Yu, and Qing Zhang. Anonymizing bipartite graph data using safe groupings. *Proc. VLDB Endow.*, 1(1):833–844, August 2008.

[44] S. Das, Textbackslashö Egecioglu, and A. El Abbadi. Anonymizing Edge-Weighted social network graphs. *Computer Science, UC Santa Barbara, Tech. Rep. CS-2009-03*, 2009.