Network Science: Exploring Small Temporal Networks

Paolo Frazzetto, Elisabeth Griesbauer & Caspar van der Woerd

IST, Network Science, a.y. 2019/2020

October 20th, 2019

1 Introduction

This report discusses an exploratory analysis over network data that was collected in the *Infectious SocioPatterns* experiment at the Science Gallery in Dublin. The goal of the original experiment was to study disease spreading in social networks. For 69 days visitors of the exhibition wore badges that measured proximity between the visitors [1].

Similarly, the goal of the current project is to study topics in temporal networks using the given data. For this first part of the project, a simplified data set was used in order to study the basic properties of the graphs and develop an initial understanding of the data. In the second part of the project more advanced analysis will be conducted which also takes into account the temporal aspect of the data. The current report discusses basic graph properties such as degree distribution, paths and clustering coefficient. The available data consists of 69 unique graphs corresponding to each day of the experiment, therefore the results have been aggregated to cancel out noise for individual days and better understand the underlying features and network concepts.

1.1 Data & Techniques

The simplified version of the dataset, provided by the authors, features graphs were nodes represent visitors and edges represent measured proximity between the respective visitors. Edge weights correspond to a simple count of the amount of 20 second intervals where two visitors were in close proximity. Such measurements were taken by means of RFID devices. Each day of the experiment corresponds to a unique graph. Note that, although the simplified data does not contain a time attribute, the effect of time is still present in the underlying structure. This is best seen in the visualizations provided by the original authors [2], were it is observable that there is typically a spine throughout the graph, which can be explained by the time aspect. Also it is notable that there are typically smaller disconnected components in the graph. Figure 1 displays the size of the graphs for the different days. For the relatively small graphs, the amount of edges scales linearly with the amount of nodes but the average degree increases for larger graphs.

On the other hand, the raw data set is made of each interaction event, i.e. an absolute time reference and the two participants identification numbers. A preliminary analysis of the full temporal process has been carried out to lay the groundwork for further developments.

Analysis of the data has been done using Python and in particular the `NetworkX` library was used for handling graph data [3]. For the time evolution plot (and eventually animation) of a single day the software *Gephi* has been used [4].

2 Graph Analysis: Topological Properties

2.1 Degree weights

The weights in the graphs are a simple counter for the amount of 20 second intervals that two visitors were in close proximity. The weighted degree of a node therefore corresponds to the total time a visitor

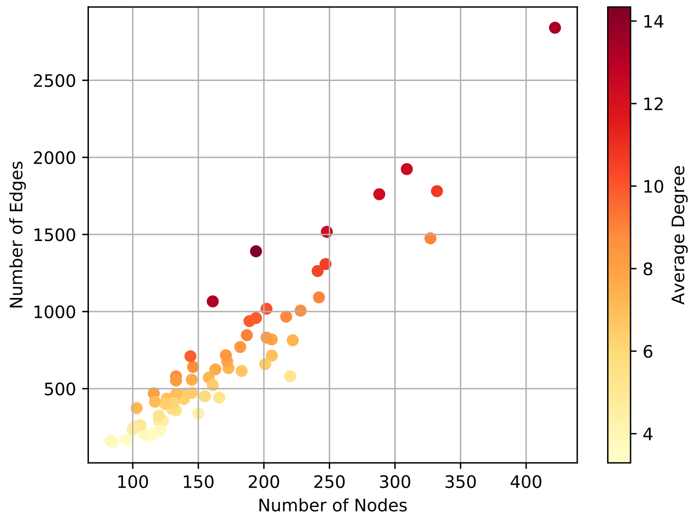

Figure 1: The amount of edges scales linearly with the amount of nodes for different days. The average degree increases significantly for larger graphs.

was in close proximity of others. If the graphs are treated as unweighted, the degree would be a count of how many other visitors a person interacted with. Naturally, it depends on the application and goal of the researcher which of these measures is more interesting. Figure 2, displays a correlation plot between weighted and unweighted degrees. A positive moderate correlation of 0.56 exists between the two. For more detailed analysis it would be interesting to examine the nodes where the proportion between weighted and unweighted degree does not follow the correlation. However, for our current research we are overall more interested in the amount of other visitors a person interacts with, therefore the unweighted edges are assumed for the remainder of this report.

2.2 Degree Distribution

The degree distributions for all 69 opening days are summarized in Figure 3. The graph is obtained by fixing degrees $k \in [k_{\max}]$ on the x-axis with $k_{\max} = 64$. Then for each degree k a box plot is created by taking into account the statistic over all the days.

Considering a degree k, the median of its daily probability densities $\rho_i(k),\ i \in [69]$, gives information about the frequency with which degree k occurs. A higher value indicates that on all 69 days having a node with degree k is more probable. We would expect that the frequency of a degree k decreases as its value increases. This is in fact true for degrees $k \geq 4$. For lower degrees the median of densities rises. One possible explanation is the nature of the experiment. On average a person spends 35 minutes at the exhibition which takes places in a museum, thus in a closed room [1, page 5]. So it is rather unlikely for a participant to have no interaction with any other person at all. We can also suggest that a certain amount of visitors attend the exhibition in company. This makes the observations plausible.

As the networks are relatively small, it is hard to determine a law for the degree distribution. However, the course of the median suggests an exponential dependence.

2.3 Path Length

In figure 4 the average shortest path is displayed against the network diameter. Note that both measures only exist for connected components, therefore the results were obtained only for the largest connected component of each graph.

The network diameter is a continuous function on the set of all shortest paths of the network. Thus one can observe a dependency between measures which in this plot appears to be linear. Observing figure 7 of [1], "the network diameter clearly defines a path connecting visitors that enter the venue at subsequent times, mirroring the longitudinal dimension of the network" [1, page 6], this is can also be

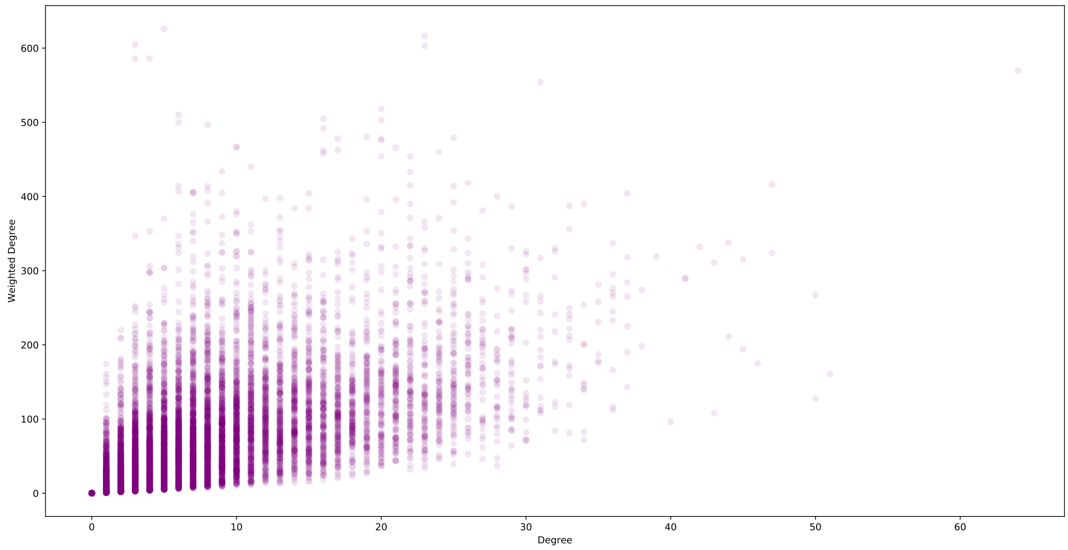

Figure 2: Correlation plot between *weighted degrees* and *unweighted degrees* of all nodes from all 69 days. Weighted degrees correspond to total count of 20 seconds intervals of proximity with other visitors. Degrees correspond to total amount of other visitors that were measured to be in close proximity of a given node. A positive moderate correlation of 0.56 exists between weighted and unweighted degrees.

observed in the visualizations [2]. "Indeed museum visitors are unlikely to interact directly with other visitors entering the venue more than one hour after them, thus preventing the aggregated network from exhibiting small-world properties." Observing the figure 7 of [1, page 7], "the network diameter clearly defines a path connecting visitors that enter the venue at subsequent times, mirroring the longitudinal dimension of the network" [1, page 6]. This longitudinal dimension of the net decreases, the more visitors are at the exhibition at the same time, thus the better frequented the exhibition is on a fixed day.

2.4 Clustering

Figure 5 shows the network clustering coefficient for the 69 daily networks. The average network clustering coefficient over all days is 0.44, as can be seen in the figure almost all of the networks have a clustering coefficient that is not more than 0.1 off of the average. Since the coefficients are quite comparable, we can assume that the dynamics of the social interactions do not change over different days and this might be relevant when dealing, for instance, with diseases spreading.

2.5 Preliminary Analysis of Temporal Data

The authors provided the full data set of the dynamic contact networks as well. Each line has the form (t, i, j), where i and j are the anonymous IDs of the persons in contact, and the interval during which this contact was active is $[t–20s, t]$. Data pre-processing was necessary to extrapolate the total amount of time of nodes' presences and interactions and, similarly to Figure [2], a visualisation of network dynamics is shown in 6. This pre-processing can be useful in future works to analyse the temporal aspect of the networks and by means of dedicated Python libraries such as `patphy` and `DyNetx`.

3 Discussion of Results

The current exploration of graph properties across the different days of the data gets us acquainted with the data and the properties of the networks. Most importantly we note that all the graphs are relatively small, which means that certain properties (e.g. degree distribution) are difficult to be explained by

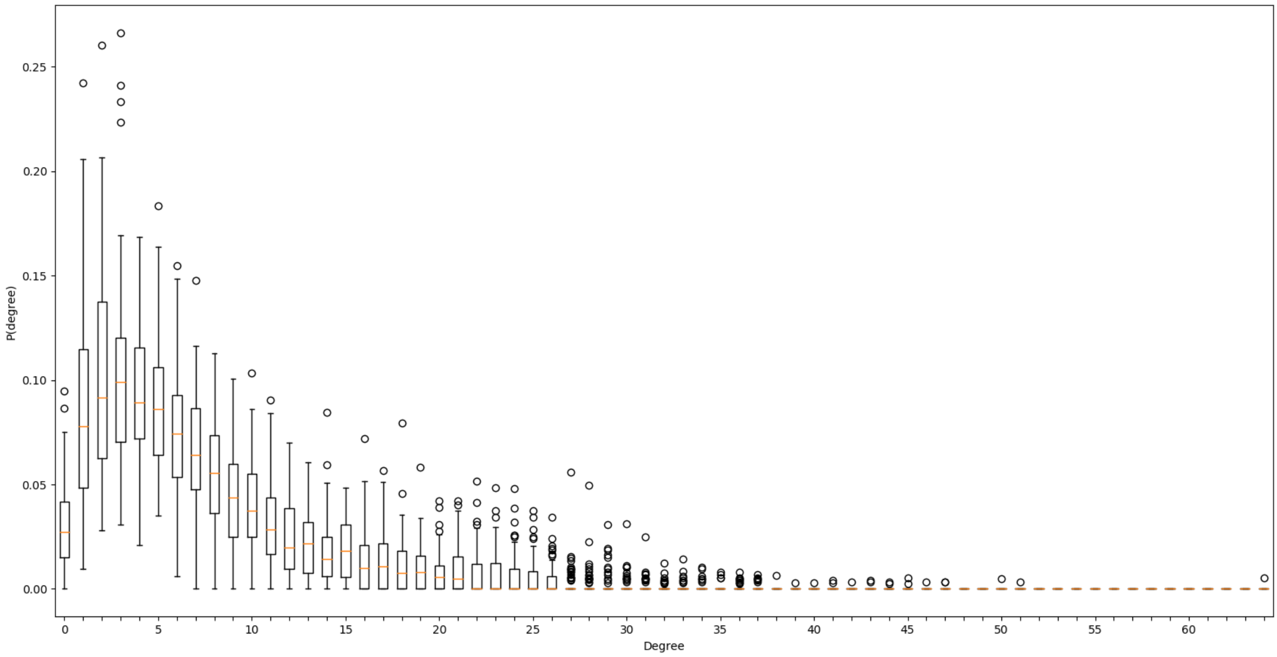

Figure 3: Aggregated degree distribution for 69 days

mathematical laws being affected by finite size effect. Furthermore, the current results suggest that across days the different graphs are quite comparable. Clustering coefficients are quite similar, the amount of edges and nodes show a linear dependence when the graph increases, as do the average shortest path and diameter lengths. For the scope of this project, this implies that we do not have to worry too much about the individual differences between different days of the data.

References

[1] 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. *Journal of theoretical biology*, 271(1):166–180, 2011.

[2] Infectious sociopatterns poster. http://www.sociopatterns.org/gallery/infectious-sociopatterns-poster/.

[3] NetworkX Developers. Networkx. https://networkx.github.io/.

[4] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

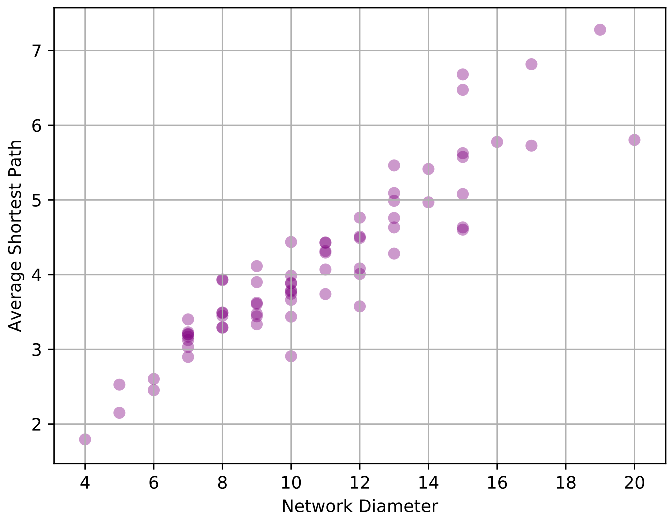

Figure 4: Average shortest path increases linearly with increasing network diameter. Both are measured in number of nodes along the path and are computed over the largest connected component of each graph.

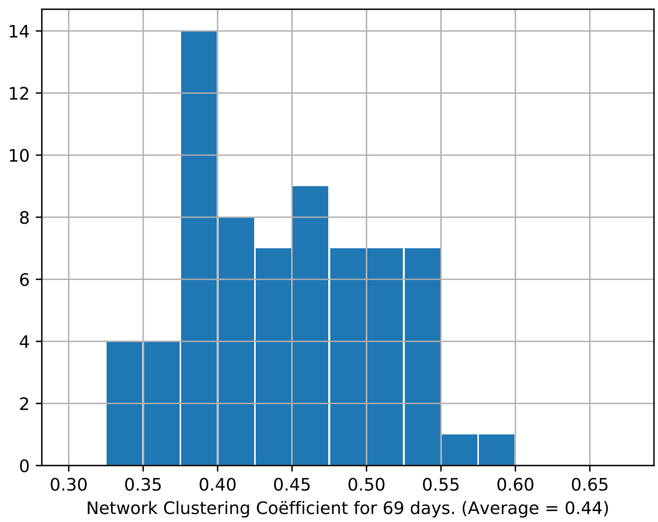

Figure 5: Network clustering coefficient for all 69 days. The values along the y-axis represent how often a clustering coefficient occurs that falls into the corresponding x-range.

Figure 6: Network dynamics for the first day of the exhibition (April 28th, 2009). Each node is colored according its degree and they are stacked clockwise according their visit time. From left to right, the network has been snapshot one hour after opening (11:00), at 14:00 and at closing time (18:00). Notice for example that there is a