Constrained Graph Resampling for Group Assessment in Human Social Networks

Nicolas Tremblay[1,⋆], Pierre Borgnat[1], Jean-François Pinton[1], Alain Barrat[2,3], Mark Nornberg[4], and Cary Forest[4]

[1] Université de Lyon, ENS Lyon, Laboratoire de Physique, CNRS UMR 5672, France
[2] Centre de Physique Théorique de Marseille, CNRS UMR 6207, Marseille, France
[3] Data Science Laboratory, ISI Foundation, Torino, Italy
[4] University of Wisconsin, Physics Department, Madison, USA

Abstract. The increasing availability of time -and space- resolved data of human activities and interactions gives insight into the study of both static and dynamic properties of human behavior. In practice, nevertheless, real-world datasets can often be considered as only one realisation of a particular event, giving rise to a key issue in social network analysis: the statistical significance of these properties. We focus in this work on features regarding groups of the networks and present a resampling - a.k.a. bootstrapping - method that enables us to add confidence intervals to such features. This in turn gives us the opportunity to compare groups' behaviors within any network. We apply this method to a new high resolution dataset of face-to-face proximity collected during two colocated scientific conferences, and it enables us to probe whether or not co-locating two conferences is an effective way of bringing together two different communities.

Keywords: Complex System, Dynamic Network Analysis, Graph Resampling, Bootstrap.

1 Introduction

High resolution experiments on face-to-face interactions between individuals in different social gatherings - such as scientific conferences, museums, schools, or hospitals - were made possible by the use of small radio sensors worn by participants, communicating with each other by bluetooth, wireless or active RFID (Radio Frequency Identification Device). These new data paved the way to many empirical investigations [3, 4, 7, 9] of human contacts, both static (existence of communities, clustering, distribution of degrees..) and dynamic (distribution of duration of contacts, of intercontacts, or of groups of different sizes..). An important issue regarding the analysis of these datasets is that each one of them can be considered as only one realisation of a particular event, it is therefore

[⋆] corresponding author. Email: firstname.lastname@ens-lyon.fr challenging to estimate confidence intervals to any of the measurable features. To this end, in the general non-network case, two methods based on constructing many random samples from the unique original data, have been widely used: the jackknife and the bootstrap methods [5]. In the case of networks, however, it is not clear how to directly transpose the classical bootstrap approach to graphs [6, 10]. In this paper, we focus on features of groups of a network, and formulate a resampling method that enables us to gain statistical significance by comparing the real data with random pseudosamples found under well-chosen constraints. The method is then applied to a dataset collected in two co-located conferences involving two distinct communities: it enables us to assess to what extent both communities mix together.

2 Resampling method for complex human contact networks

There are admittedly several ways to model a human contact network by a graph, but one can always end up with a weighted graph where each node is an individual and where the strength of the interaction between two nodes is quantified by the weight of their associated link. In the following we consider such a weighted graph as well as a group of nodes within the graph that we call X^0, whose behavior we will compare to the behavior of random groups called bootstrap samples. Let us call R^0 the group of nodes that are not in X^0. We quantify X^0's "behavior" by looking at seven observables: N^0_{XX} the total number of links within X^0, N^0_{RR} the total number of links within R^0, N^0_{XR} the total number of links connecting the two groups, T^0_{XX} the total weight of intra-X^0 links, T^0_{RR} the total weight of intra-R^0 links, T^0_{XR} the total weight of the links connecting the two groups, and Q^0_X the modularity computed for the partitioning in two groups X^0 and R^0. The modularity, in this case of a partition in two groups, is a scalar between -0.5 and 0.5 and measures how well a particular partition of the nodes separates the network into distinct communities (a value tending to 0.5 denotes two strong communities) [8]. Depending on the specific issue addressed, other observables could be considered. The backbone of the resampling protocol is the following. First, formulate a Null Hypothesis regarding the behavior of X^0. Then, compute the behavior of a large number N of groups randomly chosen within the graph called bootstrap samples (we use X as a generic notation for the bootstrap samples) for which the Null Hypothesis is true. The novelty is that each bootstrap sample is a random group *under constraint* drawn with replacement. Finally, compare the behavior of X^0 to the statistical behavior of the bootstrap samples, and decide whether or not we can reject the hypothesis. If it is rejected, a measure d is proposed to compute to what extent X^0 differs from the bootstrap samples and hence the Null Hypothesis. The comparison between different groups' behavior now boils down to the comparison of the scalar d associated to each group.

3 The data

We apply this method to a face-to-face proximity dataset collected in Salt Lake City in November 2011 during two co-located scientific conferences jointly organised by DPP (APS' Department of Plasma Physics) and GEC (Gaseous Electronics Conference) in an attempt to bring both communities – academic researchers and engineers respectively – together. In order to measure face-to-face proximity between the conference attendees wearing them, we use low power RFID tags embedded in conference badges, using the SocioPatterns sensing platform [1]. Two tags exchange packets only if they face each other (the human body acts as a shield at this frequency and power) within a distance of 1 to 1.5 meters. As soon as a tag receives a packet from another tag, it immediately uploads this information to RFID readers installed in the environment. By aggregating the five days of collected information, we obtain the overall network of contacts between the 320 participants of the experiment. Comparison with other similar experiments, and analysis of the contact patterns will be detailed in a later communication.

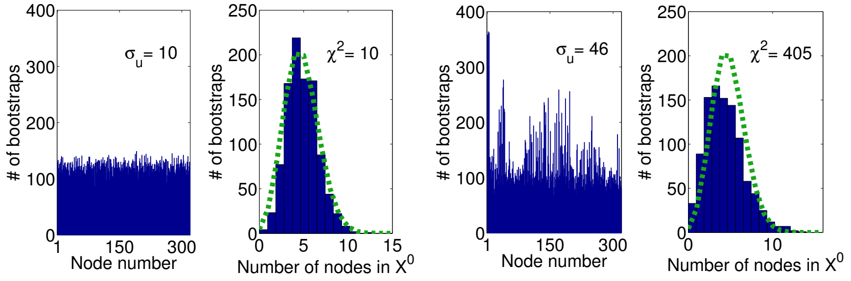
4 Is it worthwhile to co-locate both conferences?

4.1 Translating the question

In terms closer to our graph approach, we translate this question in: how well do GEC nodes mix with DPP nodes? Of course, we cannot expect GEC to mix as well as any random group of the graph: it is a community. Hence, in order to answer to the question, we apply our method to assess the difference – or similarity – between GEC's behavior with three other known communities (that will act as a benchmark): the senior researchers from DPP (SEP), the juniors from DPP (JUP), and the students from DPP (STP). The group noted X^0 in the method will alternatively be GEC, SEP, STP, or JUP. We test those four groups to the same Null Hypotheses and compare the degree with which they reject them. To show that GEC's behavior is peculiar, we look for the appropriate Null Hypothesis – if it exists – that significantly discriminates GEC from the other groups. In the following, the aggregated graph is pre-processed by deleting links that have a total time of existence inferior to 1 minute (filtering threshold under which we consider the measurement to be noise).

4.2 Same cardinal constraint for GEC

Consider the following Null Hypothesis: GEC behaves like any group of $V_X = 39$ individuals in the conference. Here, the only constraint we impose to the bootstrap samples is to have a cardinal equal to V_X. We normalize each observable Z: $z = \frac{Z - \bar{Z}^*}{\sigma_Z^*}$ where $\bar{Z}^*$ is the expected value and σ_Z^* the standard deviation in a random graph with same total number of links and same weight sequence (this is done by randomly re-allocating the weights within the ensemble of possible links). Note that $\bar{Z}^*$ and σ_Z^* depend on V_X. We choose this mode of representation for its clarity (we can plot all 7 observables on the same figure) but also because it removes the effects due to the scale of the groups allowing us to compare the results between different groups. For each normalized observable z, we define d_z the distance between the actual measured value z^{X^0} and the interval $\bar{z}^b \pm 3\ \sigma_z^b$ ($d_z = 0$ if z^{X^0} is in the interval), where $\bar{z}^b$ and σ_z^b are computed on the bootstrap samples. This interval has the meaning of an acceptance interval for the Null Hypothesis. The sum d of the distances d_z computed for each observable is the measure we use to evaluate to what extent GEC rejects the Hypothesis: the larger is d, the higher is our confidence level to reject the Hypothesis.



a)

b)

Fig. 1. Results for X^0 = GEC for a) the same cardinal constraint, b) the same cardinal and same modularity constraint. Left: histogram of the number of occurrences of each node in the bootstrap samples and its standard deviation σ_u. Right: histogram of number of X^0-nodes in a bootstrap sample with its χ^2 distance from the theoretical hypergeometric histogram (dotted line).

For all the results presented in the following, we use $N = 1000$ bootstrap samples. In Figure 1.a, we plot two histograms. The first one shows the number of times each node was chosen in a bootstrap sample. In the top right hand corner is its standard deviation σ_u: it indicates how uniformly the nodes were chosen. The second histogram shows how many nodes from GEC are in each bootstrap sample. The green dotted line represents the theoretical hypergeometric histogram computed for this same cardinal constraint. In the top right hand corner of the figure is the χ^2 distance between the observed and theoretical histograms. Each χ^2 value is computed with 10 bins that contain at least five realisations. An important point is that we do not use χ^2 for a goodness-of-fit test. In fact, we expect χ^2 to increase as soon as we impose stronger constraints on the bootstrap samples. The idea is that in the extreme case where we impose the boostraps to be exactly the GEC group, the distance d will obviously be null, σ_u will be larger than 300 and χ^2 larger than 10^{48} (the expected number of bootstrap samples having 39 GEC nodes is 10^{-48}), but we will have gained zero information. Therefore, we use χ^2 and σ_u as two control parameters of the "randomness" of the test, and make sure they stay reasonably small.

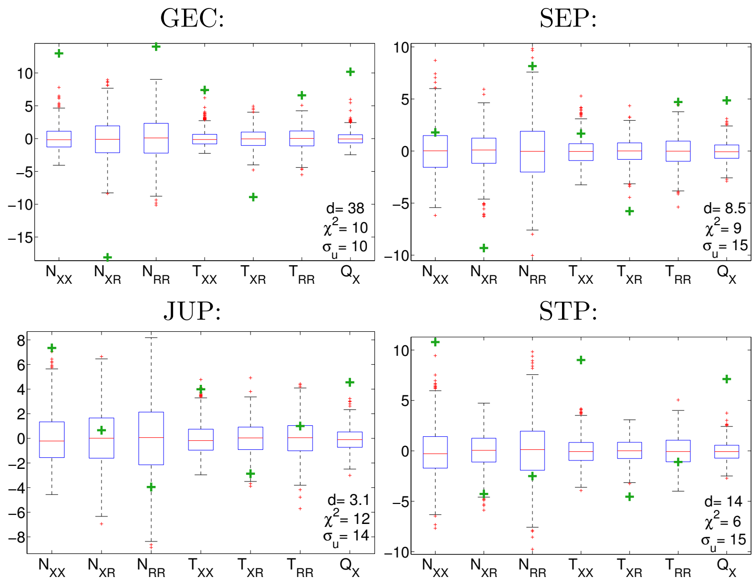


Fig. 2. Results of the same cardinal test. For each group X^0 = GEC, SEP, JUP and STP, the scalar d (bottom right hand corner of each figure) is an estimation of the distance between the statistical behavior of the bootstrap samples (boxplots) and the real data (big green crosses). χ^2 and σ_u are two control parameters of the "randomness" of the test – see text.

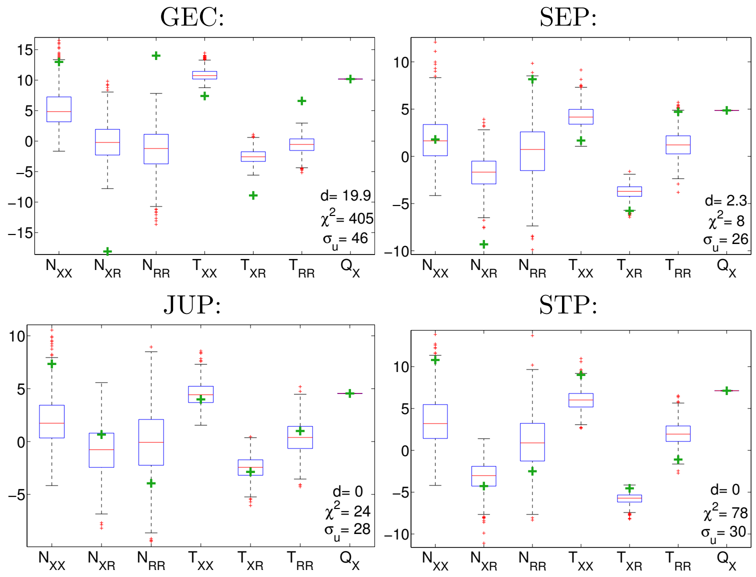


Fig. 3. Results of the same cardinal and same modularity test for the four groups.

Finally, the top left figure of Figure 2 summarizes the same cardinal constraint test for GEC: it compares the boxplots of the bootstrap samples with the measured behavior of GEC (big green crosses) and indicates d, σ_u and χ^2 in the bottom right hand corner of the figure.

4.3 Same cardinal constraint for all four groups

Figure 2 shows the results for the four groups. If the graph was random, we would have boxplots centered around zero and whiskers between – typically – -3 and 3 (corresponding to more than 99% coverage if the data was normally distributed). It is not the case (not for the whiskers) and this is an indirect proof that the graph is not random. Also, all groups have a non-null distance. This is not a surprise because those groups are known communities and behave as such: compared to the bootstrap samples, they tend to have high Q_X, N_{XX}, N_{RR}, T_{XX}, T_{RR} and low N_{XR}, T_{XR}. Interestingly, GEC's distance is clearly larger than the others: with this first naïve test, it already shows a peculiar behavior. However we can not say with statistical significance that its interaction with the rest of the conference is different from that of any specific group of people such as students or seniors (possibly also prone to discuss more with other people from their group).

4.4 Other tests

To clearly show GEC's peculiar behavior, we need to find the appropriate test – if it exists – rejected by GEC but not by the others. To this end, we need to find a compromise between strong enough constraints on the bootstrap samples to make the test more discriminative, but, as previously explained, loose enough so as to preserve the randomness of the test. In the first test, the only constraint we imposed on the boostraps was to have the same cardinal as X^0. We now refine the Null Hypothesis: X^0 behaves like any random group (with same cardinal) that has the same modularity, hence forming a community as strong as X^0. Requiring the exact same modularity is too strong a constraint and we relax it to: $Q^0_X(1-\delta) \leq Q_X \leq Q^0_X(1+\delta)$ with δ the error we tolerate. In the following, $\delta = 0.5\%$. We use a simulated annealing algorithm [2] to find such bootstrap samples and we plot the results for the four different groups in Figure 3. First, we see that the boxplots are not centered around zero anymore, they indeed need to be in accordance with a high modularity. STP and JUP's distances are null. SEP's distance is almost ten times smaller than GEC's distance: this test is a satisfying confirmation that GEC behaves differently than the other groups. We plot in Figure 1.b the two same histograms as in Figure 1.a but for bootstrap samples under these new constraints (for $X^0 = \text{GEC}$). As expected, they show a higher σ_u and χ^2, yet not so large that the randomness of the bootstrap samples would be questionable.

We also considered other kinds of constraints, for instance: keep N_{XX} constant, or keep the sum $T = 2\times T_{XX} + T_{XR}$ constant (total time of conversation of nodes in X). Results are not plotted, but the distance to the bootstrap samples is always significantly bigger for GEC than for the other groups. Furthermore, these tests are robust with respect to the filtering threshold we choose: results are similar for a filtering of 1, 3 and 5 minutes.

5 Conclusion and on-going work

We propose here a generic method to compare the behavior of different groups within a given graph. The method is inherently flexible: depending on the issue addressed in the data at hand, some observables and Null Hypotheses will be more appropriate than others. Furthermore, this method can be applied to any type of data that can be modelled by graphs. We are currently working on applying this general method to Null Hypotheses involving the dynamical behavior of groups, not only their aggregated behavior over time.

Acknowledgments

We thank the SocioPatterns collaboration [1] for providing privileged access to the SocioPatterns sensing platform that was used in collecting the contact data.

References

1. *www.sociopatterns.org.*
2. S. P. Brooks and B. J. T. Morgan. Optimization using simulated annealing. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(2):pp. 241–257, 1995.
3. C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.
4. N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.
5. B. Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Society for Industrial and Applied Mathematics Philadelphia, 1982.
6. H. Eldardiry and J. Neville. A resampling technique for relational data graphs. In *Proceedings of the 2nd SNA Workshop, 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008.
7. P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251. ACM, 2005.
8. M.E.J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.
9. M. Salathé, M. Kazandjieva, J.W. Lee, P. Levis, M.W. Feldman, and J.H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.
10. X. Ying and X. Wu. Graph generation with prescribed feature constraints. In *Proc. of the 9th SIAM Conference on Data Mining*, 2009.