From Wireless Contacts to Community Structures

Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
lastname@tik.ee.ethz.ch

***Abstract*—Human mobility and resulting contacts are driven by *intention*, *co-location*, and *social* relations between people. Based on wireless contact traces (Bluetooth, Wifi), we aim at characterizing the structure in human contacts. Instead of investigating the microscopic properties of contacts (e.g., duration and occurrence distributions), we are more interested in a macroscopic view of mobility that can more easily capture the range of human inter-relations. We hence turn to community detection. However, since these algorithms require one-dimensional tie strength metrics, we present a method to map contacts features (evolving with time) to a scalar feature value. We then analyze the outcome of the community detection by looking at inter- and intra-community ties. This provides interesting insights on the diversity of human inter-relations, which have applications to diffusion processes, for example.**

I. Introduction

The rapid proliferation of smartphones with wireless networking capabilities (Bluetooth, Wifi) creates ample opportunity for opportunistic networks where devices connect to other devices in proximity (when within radio range), “on the fly”, to exchange or spread information. This is a novel networking paradigm that is envisioned to co-exist with (and often complement) existing broadband wireless technologies (e.g. cellular, WiFi, etc.). Since actions of interest can only occur during a wireless contact, *contacts* and their statistical properties become of key importance in the design and performance evaluation of such opportunistic networks. To this end, a number of efforts have been made recently to collect relevant mobility data and analyze contact patterns; this is done either implicitly, by looking at the access points and base stations users are associated with over time in WiFi or cellular networks [1], or explicitly with experiments designed to log peer contacts (e.g. via Bluetooth) [2], [3], [4]. The majority of these traces reveal a considerable heterogeneity in contact patterns, but also significant structure and (statistical) predictability of these patterns e.g. due to time-of-day periodicity, location preference, etc. Nevertheless, the vast majority of trace analysis research in networking has focused on the *inter-contact* and *contact duration* statistics [5], [6], which are important for network performance analysis but limits mobility analysis to a microscopic view.

Recently, researchers have been looking at mobility at large-scale [1], its predictability [7], and spatial connectivity properties [8]. Human mobility and resulting contacts are actually driven by *intention*, *co-location*, and *social* relations between nodes (e.g. friends, colleagues). The latter influences someone to decide the destination (and often time) of a mobility trip; *location* on the other hand dictates the path, as well as (unknown) nodes encountered regularly at preferred/home locations (“familiar strangers”) or occasionally (“random encounters”). This creates a rather intricate contact structure that is not readily observable or usable at contact and inter-contact pattern levels. To this end, a more abstract, *macroscopic* view of mobility is needed that can more easily capture the range of node inter-relations.

In this abstract, we present a detailed study and comparison of the community structure of 3 mobility traces, namely the Haggle trace [2], the MIT Reality Mining trace [3], and the ETH trace [4]. We apply a state of the art community detection algorithm [9] to study the nature of inter-community links (e.g. bridging links vs. bridging nodes vs. community overlap, etc.), and the inter- and intra-community weight distributions in order to highlight the diversity of human relations. To our best knowledge, this is the first in depth comparative study of these properties.

In our context, nodes and contacts can be represented on a *contact graph*, where a link between two nodes indicates a measured “strong” relationship between nodes (e.g. frequent meetings, or a recent meeting [10]) through its existence (binary graph) or an edge weight (weighted graph). A variety of metrics and algorithms could then be used to characterize *node importance* on this graph, such as degree centrality, PageRank, etc., as well as to identify *similar* nodes through (implicit or explicit) *community detection*. Yet, the actual “social properties” of mobility traces, such as the modularity of communities and the distribution of inter- and intra-community weights, have not received the same amount of attention. These properties are particularly important for two reasons: *First*, they allow us to better understand the underlying structure governing human mobility and facilitate the design of improved mobility models. *Second*, they give hints on the impact of the social structure on the dynamics of diffusion processes e.g., in terms of delays but also in terms of capacity (or conductance).

The outline of this abstract is the following. In Section II, we describe the contact data used for our analysis and how we pre-process them by mapping and aggregating pair-wise contacts (i.e., different characteristics evolving over time) to a scalar value suited for community detection algorithms with weighted edges. In Section III, we analyze the outcome of the community detection algorithm. Eventually, we conclude by discussing ongoing work in Section IV.

II. Data Description

We start by describing the data used for our analysis in Section II-A. We then describe a metric of tie strength based on the principal component of contact frequency and duration (Section II-B).

A. Contact Traces

We define a *contact* as the period of time during which two devices are within radio transmission range of each other.

	MIT	INFO	ETH
Scale and context	97 campus students and staff	41 conference participants	20 lab students and staff
Period	9 months	3 days	5 days
Periodicity	300s (Bluetooth)	120s (Bluetooth)	0.5s (WiFi)
# Contacts			
Total	$100'000$	$22'459$	$23'000$
Per dev.	$1'030$	547	$1'150$

TABLE I
CONTACT TRACES CHARACTERISTICS.

A contact contains of the information about the two nodes involved, a starting time and a duration. In a opportunistic network, such a contact is an opportunity to exchange or spread information.

In order to cover a broad range of mobility scenarios with our analysis, we use different measured contact data: the MIT *Reality Mining* [3] (**MIT**), the iMotes Infocom 2005 (**INFO**) [2] and the ETH [4] (**ETH**). Their characteristics are summarized in Table I. Note that in the MIT trace, despite its long duration, a lot of short contacts were supposedly not logged due to its time granularity of 5 minutes. For our evaluation we cut the trace at both ends and used $100'000$ contacts reported between September 2004 and March 2005. Note that this time period contains holidays and semester breaks and thus still captures varying user behavior. The ETH trace contains more than $23'000$ reported contacts and is unique in terms of time granularity and reliability. Although its measurement period spans a considerably shorter time than MIT, we have on average more than 1000 reported contacts per device. This is roughly the number of contacts per device we also have for the MIT trace.

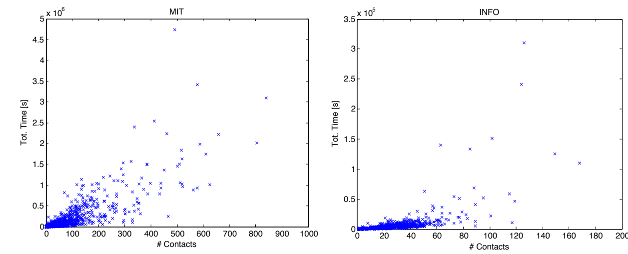
B. Tie Strength

To assess the strength of the tie between two nodes in a contact graph different metrics such as the age of last contact [11], contact frequency [12], [13] or aggregate contact duration [13] have been used (i.e., in DTN routing protocols). Here we consider two features: contact *frequency*[1] and aggregate contact *duration*. In a first step, we assign each pair of nodes $\{i, j\}$ a two-dimensional feature vector $\mathbf{z_{i,j}} = (f_{i,j}, d_{i,j})$, where $f_{i,j}$ is the number of contacts in the trace between nodes i and j, and $d_{i,j}$ is the sum of the durations of all contacts between the two nodes – both dimensions centered (zero empirical mean) and normalized to their respective standard deviation.

Figure 1 shows the scatter plots of the number of contacts vs. the total contact duration (pair-wise) for the MIT and INFO traces. They clearly show a high correlation between both features.

Since state-of-the-art community detection requires one-dimensional tie strength metrics, we transform the two-dimensional feature vector to a scalar feature value: We use the *principal component* (e.g., [14]), i.e., the direction in which the data vector Z has the largest variance the direction of the Eigenvector $\mathbf{e_1}$ with the largest corresponding Eigenvalue. We define the tie strength between i and j as the projection of $\mathbf{z_{i,j}}$ on the principal component, $w_{i,j} = \mathbf{e_1}^T \mathbf{z_{i,j}} + w_{\text{min}}$, where we add w_{min} – the smallest tie strength of all node pairs – in order to have positive tie strengths. With this metric we are able to combine the frequency and duration in a scalar value that naturally represents the heterogeneity of node pairs. We can now define the weight matrix $\mathbf{W}$ with the respective $w_{i,j}$.

[1]Note that contact age – assuming a stationary contact process – can be considered an approximation of contact frequency, therefor we do not consider it here explicitly.



(a) MIT (b) INFO

Fig. 1. Scatter-plots of number of contacts vs. total contact duration over the whole duration of the traces.

Note that with this aggregation of the contact data, we loose the timing information about contacts. We are not so much interested in the actual timing of the contacts, but rather try to capture the underlying structures that govern mobility.

The number of communities and the resulting modularity is given for each contact trace in Table II.

III. COMMUNITY STRUCTURE ANALYSIS

We will now focus on the community structure of human contacts contacts. Using the Louvain as well as Spectral community detection algorithm and the Newman modularity metric, we will first (Section III-A) assess how strongly modular contacts are. In a second step (Section III-B), we will focus on the the conductance *between* the communities, and assess how strongly communities are connected to other communities and how the conductance between them is distributed (i.e., bridging links, bridging nodes or hierarchical overlap).

A. Intra-Community Ties

In order to assess the *modularity* of a given partition of nodes to communities we compute the widely used Q function as introduced by Newman [15]. The Q function

$$Q = \frac{1}{2m} \sum_{ij} \left(w_{i,j} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j),$$

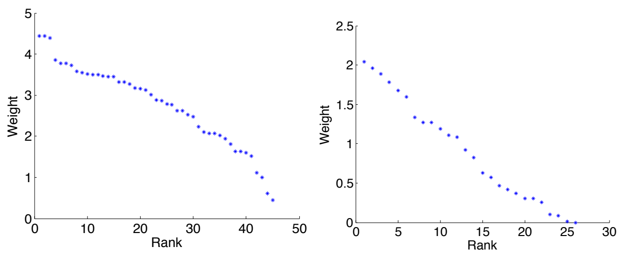
where $k_i = \sum_j w_{i,j}$ is the strength of node i and $m = \frac{1}{2} \sum_j k_j$ is the total weight in the network. c_i denotes the community of node i thus, the Kronecker delta function $\delta(c_i, c_j)$ is one if nodes i and j share the community and zero otherwise. $Q = 0$ is the expected quality of a random community assignment and [15] reports modularities of above $Q = 0.3$ for different networks (social, biological, technical, etc.) for state-of-the-art community detection algorithms[2].

In Table II we present some statistics of the trace networks' community structure as found by Louvain. A first thing to note is that the two clustering algorithms find different communities but with similar modularity. In general the modularity of the Louvain communities is slightly higher communities than the Spectral. In the rest of the paper, we will present all results for the Louvain algorithm, though, the results hold also for

[2]Note that the quality of a community assignment is a function of (i) the network, since it can be more structured or less, and (ii), the community detection algorithm, since it can find a good community assignment or not.

Trace/Model	# Comm.	Q
MIT	5	0.49
ETH	2	0.23
INFO	6	0.12

TABLE II
NUMBER OF COMMUNITIES AND MODULARITY (Q) OF CONTACT TRACES USING THE LOUVAIN ALGORITHM.



(a) ETH, Community 1 (b) MIT, Node 41

Fig. 2. Ranked Community Internal Weights (per Community and per Node).

the Spectral clustering. Second, the modularity varies broadly among the traces. We observe a strongly modular MIT trace, lower modularity in the ETH case and very low modularity in the INFO case. Similar values for other community detection algorithms (K-Clique and Newman), different traces and other strength metric (total contact duration) have already been reported in [13], thus we confirm these findings as a first result.

To find out more about the insides of communities we look at the distribution of intra-community weight. Figure 2 shows some typical representatives of community-internal tie strengths, ranked over all node pairs of a community, as well as per node. We observe that the weights are strongly skewed. A community can thus not be thought of as a homogeneous group of strongly connected nodes (like a mesh). Instead, there is strong heterogeneity even within a community. This observation is consistent throughout all traces and all communities (only few are shown in Figure 2 due to space limitations).

B. Inter-Community Ties

Comm. Index	1	2	3	4	5
1	**20.5%**	1.0%	0.5%	0.1%	0.01%
2	1.0%	**31.8%**	4.2%	2.9%	0.2%
3	0.5%	4.2%	**13.4%**	2.7%	0.2%
4	0.1%	2.9%	2.7%	**8.9%**	0.1%
5	0.01%	0.2%	0.2%	0.1%	**2.1%**

TABLE III
PERCENTAGES OF TOTAL WEIGHT WITHIN AND BETWEEN COMMUNITIES (MIT TRACE). ALL WEIGHTS SUM TO 100% AND INTER-COMMUNITY WEIGHTS ARE HALVES BETWEEN TIED COMMUNITIES.

We now change our focus on the interface *between* the communities. Table III shows an example matrix for the MIT trace of how the total weight in the network is distributed within the communities and between the communities. In the matrix we see that the inter-connections of communities are weak in many cases. For instance, communities 1 and 2 together contain more than 50% of the weights and 50% of the nodes. However, between them there is only 1% of the weight.

IV. DISCUSSION AND CONCLUSIONS

The results presented herein are preliminary investigations of using community detection algorithms to highlight the community structure of contact traces. Actually, it does not only matter how much of the weight falls between two communities, but also how this weight is distributed. Thus, we are currently aiming at identifying the type of interface as either (i) bridging links (people linked to one specific person in another community), (ii) bridging nodes (people part of two communities i.e., overlap), or (iii) hierarchical communities. We characterize these three types in the following.

Note that certain community detection algorithms inherently identify some of these interfaces. For instance the K-Clique algorithm [16] allows nodes to be in more than one community and thus identifies bridging nodes. Similarly, a class of algorithms such as Newman Girvan [17] is based on a hierarchical tree (dendrogram) and thus inherently identifies hierarchies. However, neither of them provides a distinction between all the three types of inter-connection. We are hence currently combining the peculiar features of existing algorithms at once.

REFERENCES

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.
[2] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Pocket switched networks and human mobility in conference environments," in *WDTN*, 2005.
[3] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, May 2006.
[4] V. Lenders, J. Wagner, and M. May, "Measurements from an 802.11b mobile ad hoc network," in *EXPONWIRELESS*, 2006.
[5] V. Conan, J. Leguay, and T. Friedman, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *ACM Autonomics*, October 2007.
[6] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic, "Power law and exponential decay of inter contact times between mobile devices," in *MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. New York, NY, USA: ACM, 2007, pp. 183–194.
[7] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, February 2010.
[8] P. Wang and M. C. Gonzales, "Understanding spatial connectivity of individuals with non-uniform population density," *Philosophical Transactions of The Royal Society A*, vol. 367, pp. 3321–3329, August 2009.
[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J.STAT.MECH.*, 2008.
[10] T. Hossmann, T. Spyropoulos, and F. Legendre, "Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing," in *Infocom 2010*. IEEE, March 2010.
[11] H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli, "Age matters: efficient route discovery in mobile ad hoc networks using encounter ages," in *MobiHoc*, 2003.
[12] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," *SIGMOBILE Mobile Computer Communications Revue*, July 2003.
[13] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay tolerant networks," in *MobiHoc*, May 2008.
[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.
[15] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, June 2006.
[16] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, June 2005.
[17] M. E. J. Newman, "Analysis of weighted networks," 2004.