

Multiple-Membership Communities Detection in Mobile Networks

Nikolai Nefedov
Nokia Research Center, Switzerland;
ISI Lab, ETH Zurich, Switzerland
nikolai.nefedov@nokia.com

ABSTRACT

This paper addresses a problem of overlapped communities detection and its applications in mobile networks. The suggested method is based on "soft" communities detection using modularity maximization and may be used to improve efficiency of mobile advertisements and to make recommendations and missing links predictions. The method is applied for analysis of mobile users data collected at Nokia data collection campaign. In particular, we outline possible applications of soft community detection in single and multi-layer graphs.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering;
G.2.2 [Graph Theory]: Graphs algorithms

General Terms

Algorithms, Design

Keywords

community detection, recommendations, multi-layer graphs

1. INTRODUCTION

The growing spread of smart phones equipped with various sensors allows to collect real world data on human proximity and capture social interactions. The rich mobile sensing data could help to understand complex patterns of human activities and to develop various mobile applications and services on-demand for end-users. One of the first steps in analysis of mobile datasets is to find underlying structure of the network and its dynamics by clustering data according to some similarity measures.

Classification and clustering (finding groups of similar elements in data) are well-known problems which arise in many fields of sciences, e.g., [1][2][3]. In cases when objects are characterized by vectors of attributes, a number of

efficient algorithms to find groups (clusters) of similar objects based on a metric between the attribute vectors are developed. On the other hand, if data are given in the relational format (causality or dependency relations), e.g., as a network consisting of N nodes and E edges representing some relation between the nodes, then the problem of finding similar elements corresponds to detection of communities, i.e., groups of nodes which are interconnected more densely among themselves than with the rest of the network.

The growing interest to the problem of community detection was triggered by introduction of new clustering measure called modularity [4][5]. Currently a number of different algorithms for modularity maximization are proposed (e.g., see [8]-[13] and references within). However, most of these methods address network partitions into disjoint communities. On the other hand, in practice communities are often overlapping. It is especially visible in social networks, where people are involved in different groups, depending on professional activities, family status, hobbies, and etc. Furthermore, social interactions are reflected in multiple dimensions, such as users activity, local proximity, their locations and etc. These multi-dimensional traces may be presented as multi-layer graphs. It raises a problem of overlapping communities detection in single or multi-layer graphs, which we address by introduction of soft communities detection combined with random walk. In particular, we outline possible applications of soft community detection in single and multi-layer graphs built from mobile datasets.

This paper is organized as follows: Section 2 presents an outline of community detection based on modularity maximization using fast greedy search and its extension based on dynamical properties of random walk. Soft community detection for networks with overlapping and its applications are addressed in Section 3. In particular, we suggest friend recommendations in social networks, where new link recommendations are made as intra- and inter-clique communities completion and recommendations are prioritized according to similarity measures modified to include multiple-communities membership. Some examples of community detection applied for selected datasets from Nokia data collection campaign are presented in Section 4. Here we consider single and multi-layer graphs to improve community detection accuracy and present examples of friend recommendations in intra- and inter-communities.

2. COMMUNITY DETECTION

2.1 Modularity Maximization

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS '11, May 25-27, 2011 Sogndal, Norway

Copyright 2011 ACM 978-1-4503-0148-0/11/05 ...\$10.00.

Recently Newman et al [4][5] introduced a new approach called modularity which is defined as a number connections within a group compared to the expected number of such connections in an equivalent null model (e.g., in an equivalent random graph). In this case the modularity of a partition P may be written as

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(c_i, c_j) \quad (1)$$

where c_i is i -th community., A_{ij} are elements of graph adjacency matrix; d_i is degree of i -th node, $d_i = \sum_j A_{ij}$; m is total number of links $m = \sum_i d_i/2$; P_{ij} is a probability that nodes i and j are connected; in random graph $P_{ij} = d_i d_j / 2m$.

By construction $|Q| < 1$ and $Q = 0$ means that the network under study is equivalent to the used null model (an equivalent random graph). Case of $Q > 0$ indicates a presence of a community structure, i.e., more links remain within communities than would be expected in the equivalent random graph. Hence, a network partition which maximizes modularity may be used to locate communities. This maximization is NP-hard and a number of suboptimal algorithms are suggested.

In the following we use greedy search algorithm [6] extended with random walk approach which gives a reasonable trade-off between accuracy of community detection and scalability.

Greedy Search Algorithm

Input: a graph as $N \times N$ matrix.

Initialize each node i as a community c_i with modularity $Q(i) = -\left(\frac{d_i}{2m}\right)^2$.

Repeat while there is an increase in modularity:
for all nodes i **do**:

remove i from its community c_i ;

insert i sequentially in neighboring communities c_j for all j with $A_{ij} > 0$;

calculate modularity $Q(c_j^{(i \rightarrow c_j)})$;

select a move (if any) of i -th node to community c_j^* with max modularity

$$Q(c_j^{(i \rightarrow c_j^*)}) = \max_{j \in N(i)} Q(c_j^{(i \rightarrow c_j)});$$

Stop when (a local) maximum is reached.

2.2 Dynamic Formulation of Modularity

It is well-known that a network topology affects a system dynamics, it allows us to use the system dynamics to identify topology [16]. Let's assume that for an undirected network there exist a partition $\mathcal{P}$ with communities $c_k \in \mathcal{P}$, $k = 1, \dots, M$. The probability that initially, at t_0 , a random walker belongs to a community c_k is $\Pr(c_k, t_0) = \sum_{j \in c_k} d_j / 2m$. Probability that a random walker, which was initially in c_k , will stay in the same community at the next step $t_0 + 1$ is given by

$$\Pr(c_k, t_0, t_0 + 1) = \sum_{j \in c_k} \sum_{i \in c_k} \left(\frac{A_{ij}}{d_j} \right) \left(\frac{d_j}{2m} \right). \quad (2)$$

Under assumption that dynamics is ergodic, the memory of the initial conditions is lost at infinity, hence $\Pr(c_k, t_0, \infty)$ is equal to the probability for two independent walkers to be in c_k ,

$$\Pr(c_k, t_0, \infty) = \left(\sum_{i \in c_k} \frac{d_i}{2m} \right) \left(\sum_{j \in c_k} \frac{d_j}{2m} \right). \quad (3)$$

Combining (2) and (3) we may write

$$\begin{aligned} \sum_{c_k \in \mathcal{P}} (\Pr(c_k, t_0, t_0 + 1) - \Pr(c_k, t_0, \infty)) \\ = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(c_i, c_j) = Q \end{aligned} \quad (4)$$

In general case one may define stability of $\mathcal{P}$ partition [15][16]

$$\begin{aligned} R_{\mathcal{P}}(t) &= \sum_{c_k \in \mathcal{P}} \Pr(c_k, t_0, t_0 + t) - \Pr(c_k, t_0, \infty) \\ &= \sum_{c_k \in \mathcal{P}} \sum_{i,j \in c_k} \left(\left(e^{t(B-I)} \right)_{ij} \frac{d_j}{2m} - \frac{d_i d_j}{4m^2} \right), \text{ where } B_{ij} = \frac{A_{ij}}{d_j}. \end{aligned}$$

Furthermore, as shown in [15], we may define time-varying modularity $Q(t)$ by linear terms in time expansion for $R(t)$ at $t \approx 0$,

$$R(t) \approx (1-t) \cdot R(0) + t \cdot Q = Q(t)$$

which finally gives

$$Q(t) = (1-t) + \sum_{c_k \in \mathcal{P}} \sum_{i,j \in c_k} \left(\frac{A_{ij}}{2m} t - \frac{d_i d_j}{4m^2} \right) \quad (5)$$

In the following we apply time-dependent modularity maximization (5) using greedy search to find hierarchical structures in networks beyond modularity maximization Q_{max} (1). This approach is useful in cases where maximization of (1) results in very fragmental structure with a large number of communities.

3. OVERLAPPING COMMUNITIES

3.1 Multi-Membership

In social networks people belong to several overlapping communities depending on their families, occupations, hobbies, etc. As the result, users (presented by nodes) may have different levels of membership in different communities. This fact motivated us to consider multi-community membership as edge-weights to different communities and partition edges instead of clustering nodes.

As an example, we can measure membership $g_j(k)$ of node k in j -th community as a number of links (or its weight for a weighted graph) between the k -th node and other nodes within the same community. Then, for each node k we assign a vector $\mathbf{g}(k) = [g_1(k), g_2(k), \dots, g_M(k)]$, $k = \{1, \dots, N\}$ which presents the node membership (or participation) in all detected communities $\{c_1, \dots, c_M\}$. In the following we refer $\mathbf{g}(k)$ as soft community decision for node k .

To illustrate the approach, overlapping communities derived from benchmark karate club social network [7] and membership distributions for selected nodes are depicted at Fig.1 and Fig.2, respectively. Modularity maximization here reveals 4 communities shown by different colors. However,

multi-communities membership results in overlapping communities illustrated by overlapping ovals (Fig.1). For example, according to modality maximization, the node 1 belongs to community c_2 , but it also has links to all other communities indicated by blue bars at Fig.2.

Participation of different nodes in selected communities is presented at Fig.3 and Fig.4. These graphs show that even if a node is assigned by some community detection algorithm to a certain community, it still may have significant membership in other communities. This multi-communities membership is one of the reasons why different algorithms disagree on communities partitions. In practice, e.g., in targeted advertisements, due to the "hard" decision in community detection, some users may be missed even if they are strongly related to the targeted group. For example, user '29' is assigned to c_3 (Fig.1), but it also has equally strong memberships in c_2 and c_4 (Fig.3). Using soft community detection user '29' can also be qualified for advertisements targeted to c_2 or c_4 .

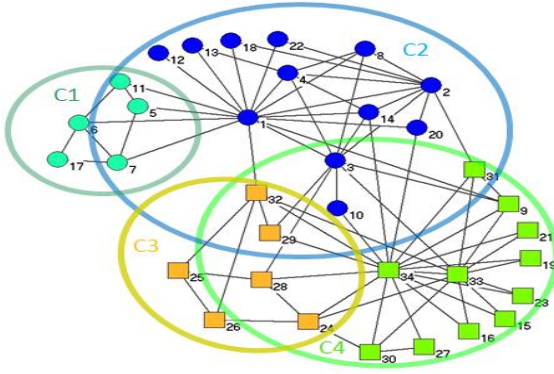


Figure 1: Overlapping communities in karate club.

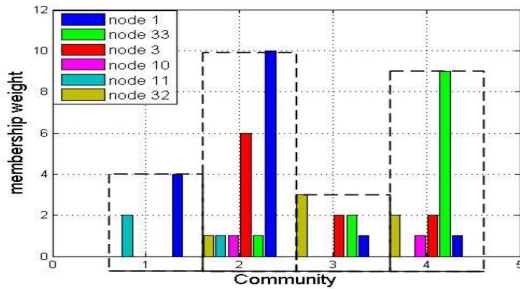


Figure 2: Membership weight distribution for selected users in karate club social network.

3.2 Applications to Recommendation Systems

In online social networks social new links recommendations may be seen as an attractive service. Recently Facebook and LinkedIn introduced a service "People You May Know", which recommends new connections using the friend-of-a-friend (FoF) approach. However, in large networks FoF approach may create a long and often not relevant list of recommendations, which is difficult (and also computationally expensive, in particular in mobile solutions) to navigate. Furthermore, in mobile social networks (e.g., Nokia portal

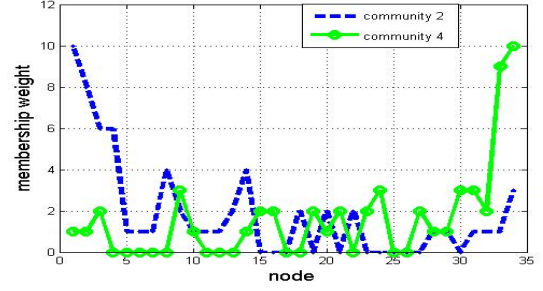


Figure 3: Karate club: participation of users in communities c_2, c_4 .

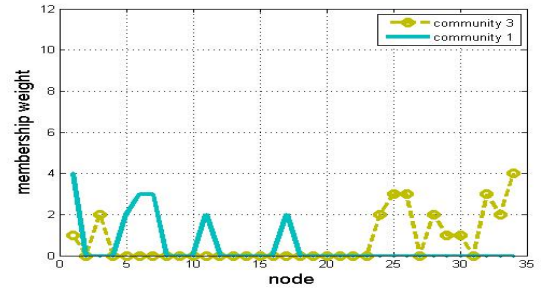


Figure 4: Karate club: participation of users in communities c_1, c_3 .

Ovi Store) these kinds of recommendations are even more complicated because users' affiliations to different groups (and even its number) are not known. Hence, before making recommendations, communities are to be detected first.

Recommendations as Communities Completion

In this paper we suggest to use soft communities detection for FoF recommendations as follows: (i) detect communities; (ii) calculate membership in different communities for each node; (iii) make new recommendations as communities completion following the rules below; (vi) use multiple-membership to prioritize recommendations. In particular, to make prioritization we introduce several similarity measures outlined below.

To make new link recommendations we suggest the following rules:

- each new link creates at least one new clique (FoF concept);
- complete cliques within the same community (intra-cliques) using the FoF concept;
- complete cliques towards to the fully-connected own community if there is no FoF links;
- complete inter-cliques (where nodes belongs to different communities);
- prioritize intra-clique and inter-clique links completion according to some measure based on multi-membership.

Modified topology-based predictors

Let's define sets of neighbors of node k , which are inside and outside of community c_i as $\Gamma_i(k) = \{\Gamma(k) \in c_i\}$. This

allows us to introduce a set of similarity measures by modifying topology-based base-line predictors to take into account multiple-membership overlapping communities.

As an example, for the intra-clique completion we may associate a quality of missing link prediction (or recommendation) between nodes k and n within c_i community with modified base-line predictors scores as follows:

- Preferential attachment: $S_{PA}^{(i,i)}(k, n) = |\Gamma_i(k)| \cdot |\Gamma_i(n)|$;
- Jaccards score: $S_{JC}^{(i,i)}(k, n) = |\Gamma_i(k) \cap \Gamma_i(n)| / |\Gamma_i(k) \cup \Gamma_i(n)|$;
- Adamic/Adar score: $S_{AA}^{(i,i)}(k, n) = \sum_{z \in \Gamma_i(k) \cap \Gamma_i(n)} (\log |\Gamma(z)|)^{-1}$;
- Katz score (intra-community):

$$S_{KC}^{(i,i)}(k, n) = \sum_{l=1}^{\infty} \beta^l |\text{path}_i(k, n)^{(l)}| = \left\{ (I - \beta A^{(i)})^{-1} - I \right\}_{(k,n)}$$

where $|\text{path}_i(k, n)^{(l)}|$ is number of all paths of length- l from k to n within c_i ; β is a dumping parameter, $0 < \beta < 1$; I is the identity matrix, $A^{(i)}$ is the (weighted) adjacency matrix of community c_i .

Another example is a community connectivity measure, $S_{CC}^{(i,i)}(k, n) \sim \lambda_2(L)$, which characterizes a convergence speed of opinions to consensus within a community c_i when a link between nodes k and n is added inside the community; here $\lambda_2(L)$ is the 2nd smallest eigenvalue of graph Laplacian $L = D - A$ (or the normalized Laplacian for weighted graphs) of community c_i .

The measures above consider communities as isolated ones and may be used as an approximation based on disjoint communities partitions. To take into account both intra- and inter-community links we use multi-community membership for nodes, $g_i(k)$. In general, for nodes $k \in c_i$ and $n \in c_j$, the inter-community relations may be asymmetric, $g_j(k) \neq g_i(n)$. In the case of undirected graphs we may use averaging and modify the base-line predictors $S(k, n)$ as

$$S^{(i,j)}(k, n) = \frac{g_j(k) + g_i(n)}{2m} S(k, n) \quad (6)$$

For example, modified Jaccards and Katz scores which take into account multi-communities membership are defined as

$$S_{JC}^{(i,j)}(k, n) = \frac{g_j(k) + g_i(n)}{2m} \frac{|\Gamma(k) \cap \Gamma(n)|}{|\Gamma(k) \cup \Gamma(n)|} \quad (7)$$

$$S_{KC}^{(i,j)}(k, n) = \frac{g_j(k) + g_i(n)}{2m} \left\{ (I - \beta A^{(C_{n,k})})^{-1} - I \right\}_{(k,n)} \quad (8)$$

where $k \in c_i, n \in c_j$; $A^{(C_{n,k})}$ is an adjacency matrix formed by all communities relevant to nodes n and k .

Recommendations also may be made in the probabilistic way, e.g., to be picked up from distributions formed by modified prediction scores.

4. APPLICATIONS FOR MOBILE DATA

4.1 Nokia mobile data collection

To analyze mobile users behavior and study underlying social structure Nokia Research Center/Lausanne organized mobile data collection campaign at EPFL campus [17]. Rich-content datasets (including data from mobile sensors, call-logs, bluetooth (BT) and WLAN proximity, GPS coordinates, information on mobile and applications usage and etc)

are collected from about 200 participants for the period from June 2009 till October 2010. Besides collected data, several surveys before and after the campaign have been conducted to profile participants and to form a basis for the ground truth.

In this section we consider social affinity graphs constructed from call-logs, GPS locations and users proximity. Fig.5 shows a weighted aggregated graph of SMS and call-logs connections derived from corresponding datasets. This graph depicts connections among 136 users, which indicates that about 73% of participants are socially connected within the data collection campaign. To find communities in this network we first run the modularity maximization algorithm, which identifies 14 communities after the 3d iteration (Fig.5).

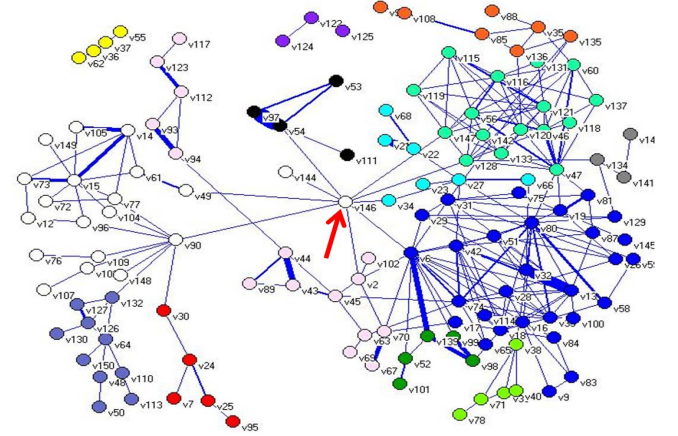


Figure 5: Community detection based on SMS and call-logs: communities are coded by colors.

Note that (local) max modularity is reached at 14 communities and to get the higher hierarchical levels one could aggregate communities into single nodes. However, this procedure would result in a loss of underlying structure. In particular, hierarchical community detection with nested communities structure poses additional constrain on maximization process which may lead to incorrect classification at higher layers. For example, after 3d iteration the node 'v146' shown by arrow at Fig.5 belongs (correctly) to community shown by white circles (3 intra-community edges and single edges to other 6 communities). After agglomeration, the 'v146' will be assigned to the community shown by white circles on the left side on the graph. However, it is easy to notice that as single communities on the right are merged, node 'v146' should be re-assigned to another community on right side of the network. Dynamical formulation of modularity with random walk allows different (not necessarily nested) allocations of nodes at different granularity (resolution) levels and helps to resolve this problem.

4.2 Applications for Multi-layer graphs

Besides call-logs, social affinity of participants may also be derived from local proximity of users (BT and WLAN layers) and their location information (GPS). In this case soft communities detection may be extended to include multiple graph layers. In particular, we found that users' profiles may significantly vary across the layers. For example, a user

may have dense BT connections with a multiple communities participation, while his call phone activities may be rather limited. Multi-layer communities analysis is a rich field of research, below we show some preliminary results.

To make verification of detected communities we selected a subset of 136 users with known email affiliations as the ground truth. In our case we got 8 groups for emails. To get the same number of communities in social affinity graphs, we used random walk (5) to get more coarse resolution than provided by modularity maximization (1).

Fig 6 depicts communities (color coded) derived from the phone-calls graph. Single nodes indicate users which did not make phone calls to participants of the campaign. Communities derived from the BT-proximity (combined BT and phone calls) graph mapped on phone-call graph are shown at Fig.7 (Fig.8). As expected, multi-layers graphs help us to classify users based on the additional information found in other layers. For example, users without phone calls (Fig.6) are assigned to communities based on BT proximity (Fig.7). To evaluate accuracy of community detection we used the normalized mutual information (NMI) score, purity test and Rand index (RI) [18]. Results are summarized in Table1. As one can see, combining information from different graph layers improves reliability of communities detection.

As expected, different graph layers have different relevance to the email affiliations and do not have fully overlapped community structures. In particular, proximity seems to be more relevant to professional relations indicated by email affiliations, while phone calls seem to reflect more friendship and family relations. However, the detected structures are still rather close to each other (cf. columns in Table 1) reflecting underlying social affinity.

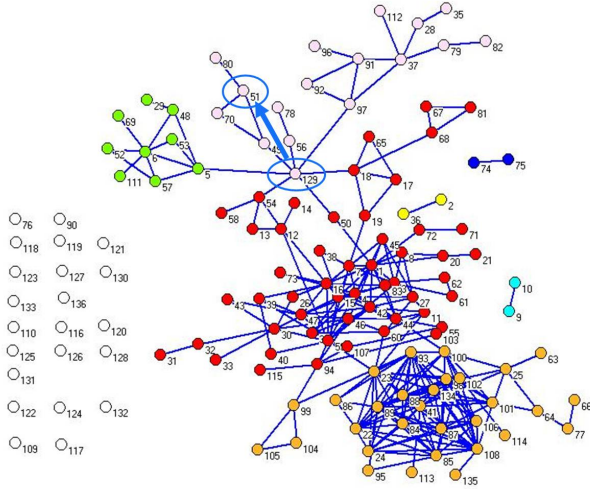


Figure 6: Community detection using random walk in phone-calls network.

4.3 Application for Recommendation Systems

As discussed in Section 3, one of applications of soft communities detection may be seen in recommendation systems. To illustrate the approach we select the user 129 (marked by oval) from phone-calls network at Fig.6 and calculate

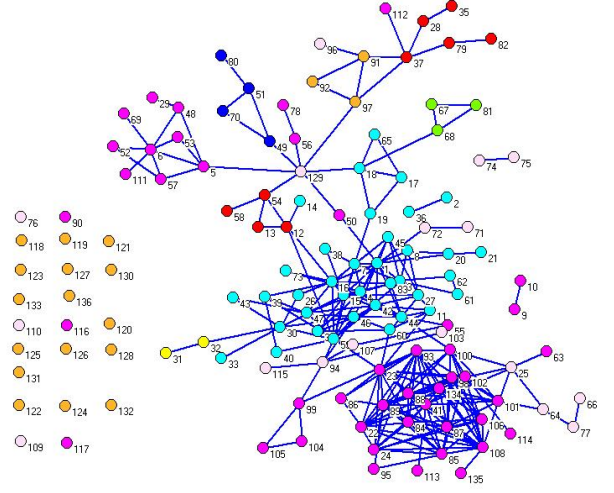


Figure 7: Community detection based on BT proximity mapped on phone-calls network.

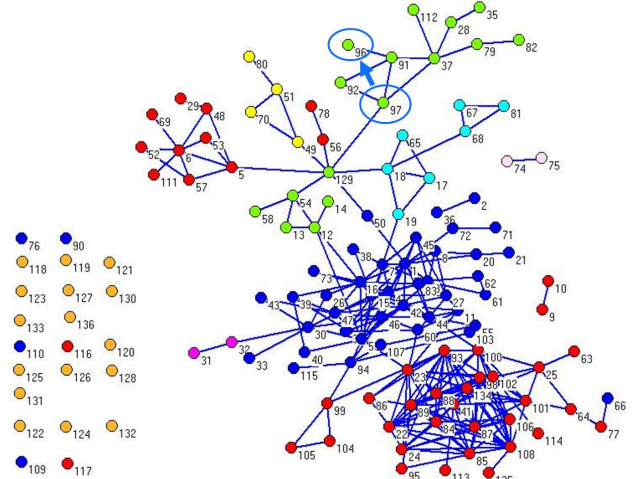


Figure 8: Community detection based on combined BT & phone-calls mapped on phone-calls network.

proposed prediction scores to prediction prioritization. The suggested FoF intra-link recommendations for two predictors (Katz predictor and convergence speed to consensus, $S_{CC}^{(i,i)}(k,n)$) are shown in Table 2. Here we list new links for user 129 which create at least one clique within its community (shown by pink color at Fig.6) and normalized prediction scores. Depending on a purpose of recommendation we may select different prediction criteria. Since new links change topology, which in turn affects dynamical properties of the network, the recommendations may be seen as a distributed control driving the network evolution. Convergence of opinions to a consensus within communities in many cases is important target in social science, so we selected it for ex-

Table 1: Community detection evaluation

	NMI	Purity	RI
Phone calls	0.262	0.434	0.698
BT proximity	0.307	0.456	0.720
GPS	0.313	0.471	0.704
Phone + BT	0.342	0.427	0.783

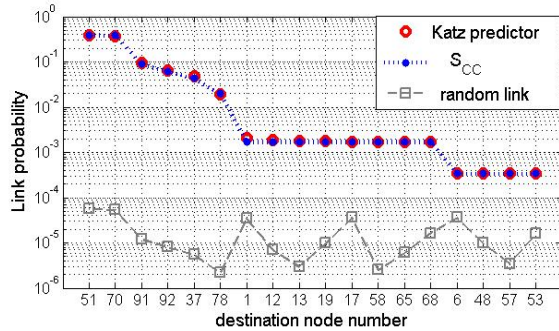
Table 2: FoF recommendations for user 129 within its own community according to different similarity measures for phone-calls network (Fig.6)

source	destination	$S_{KC}^{(i,i)}(s,d), \%$	$S_{CC}^{(i,i)}(s,d), \%$
129	51	10.5	22.6
129	78	11.1	16.3
129	91	47.1	15.4
129	70	11.3	15.3
129	92	9.6	15.3
129	37	10.5	15.1

amples below. The best intra-community recommendation for phone-calls network according to $S_{CC}^{(i)}(k,n)$ is shown by the arrow at Fig.6.

Prediction scores for cases with intra- and inter-communities links are depicted at Fig.9, here the scores are normalized as probabilities and sorted according to its priority; corresponding random links probabilities are shown as the reference. Note that the highest priority recommendation link {129,51} is the same as in the intra-community only recommendation (cf. Table 2). However, the presence of inter-communities links modifies priorities of other recommendations. To make verification we compare the predicted links at the phone-call network with links observed for user 129 at the BT proximity layer. This comparison shows a good fit: 16 out of 18 predicted links are found at the BT layer.

In general, the selection of topology-based recommendation criteria and their verifications are the open problems. Currently we are planing experiments to evaluate different recommendation criteria and its rate of acceptance.

**Figure 9: FoF recommendations for user 129 and its priorities over all relevant communities.**

5. CONCLUSIONS

In this paper we presented the method of soft community detection in large networks and its possible applications. The suggested method may be used to improve marketing efficiency by including users which are strongly relevant to the

targeted group, but are not detected by the standard community detection methods. Also, soft community detection may be applied for recommendation systems to make new link recommendations or to make predictions on missing (or unobservable) relations social networks. This method is applied for analysis of datasets recorded during Nokia mobile-data collection campaign to derive community structures in multi-layer graphs and to make new link recommendations.

6. REFERENCES

- [1] S.Wasserman and K.Faust, Social Network Analysis. Cambridge University Press, Cambridge (1994).
- [2] R.Albert and A.-L.Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47-97 (2002).
- [3] G.W.Flake, S.R.Lawrence, C.L.Giles, and F.M.Coetzee, Self-organization and identification of Web communities. *IEEE Computer* **35**, 66-71 (2002).
- [4] M.Girvan and M.E.J.Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821-7826 (2002).
- [5] M.E.J.Newman and M.Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).
- [6] M.E.J.Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E* **69**, 066133 (2004).
- [7] W.W.Zachary, An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33**, 452-473 (1977).
- [8] F.Wu and B.A.Huberman, Finding communities in linear time: A physics approach. *Eur. Phys. J. B* **38**, 331-338 (2004).
- [9] F.Radicchi, C.Castellano, F.Cecconi, V.Loreto, and D.Parisi, Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101**, 2658-2663 (2004).
- [10] A.Capocchia, V.D.P.Servedio, G.Caldarella, F.Colaorib, Detecting communities in large networks, *Physica A* **352** 669-676 (2005).
- [11] S.Fortunato, C.Castellano, Community Structure in Graphs, arXiv:0712.2716v1.
- [12] M.A.Porter, J.-P. Onnela, and P.J.Mucha, Communities in Networks arXiv:0902.3788.
- [13] S. Fortunato, Community detection in graphs, *Physics Reports* **486**, 75-174 (2010).
- [14] P.Schuetz and A.Caflisch, Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement *Phys. Rev. E* **77**, 046112 (2008).
- [15] T.S. Evans and R. Lambiotte, Line Graphs, Link Partitions and Overlapping Communities, *Phys.Rev.E* **80** 016105 (2009).
- [16] R. Lambiotte et al, Laplacian Dynamics and Multiscale Modular Structure in Networks, arXiv:0812.1770v3.
- [17] N.Kiukkonen et al, Towards Rich Mobile Phone Datasets: Lausanne Data Collection Campaign *Proc. ACM Int. Conf. Pervasive Services*, Berlin, 2010.
- [18] C.D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.