An Unsupervised Learning Approach to Social Circles Detection in Ego Bluetooth Proximity Network

Jiangchuan Zheng[♮], **Lionel M. Ni**[♮,♯]
[♮]Department of Computer Science and Engineering
[♯]Guangzhou HKUST Fok Ying Tung Research Institute
[♮]Hong Kong University of Science and Technology
{jczheng,ni}@cse.ust.hk

ABSTRACT
Understanding a user's social interactions in the physical world proves important in building context-aware ubiquitous applications. A good way towards that objective is to categorize people to whom a user is socially related into what we call as *social circles*. In this note, we propose a novel unsupervised approach that learns from the Bluetooth (BT) sensed data recording one's dynamic proximity relations with others to identify her social circles, each of which is formed along a semantically coherent aspect. For each circle we learn its members as well as the temporal dimensions along which it is formed. Our method is innovative in that it well overcomes data sparsity by information sharing, and allows for circle overlaps which is common in reality. Experiments on real data demonstrate the effectiveness of our method, and also show the potentials of relational mobile data in sensing personal behaviors beyond personal data.

Author Keywords
Bluetooth Sensing; Social Circle Learning; Human Behavior Analysis; Collaborative Filtering

ACM Classification Keywords
I.5 Pattern Recognition; H.5.2 User Interfaces

INTRODUCTION AND RELATED WORK

The proliferation of smartphone-generated sensor data in recent years offers a unique opportunity to model human behaviors using computational methods. In the past, a large body of work has leveraged data types that are personal by nature such as GPS traces and application usages to uncover the laws and patterns in individual behaviors including the study of the predictability in human mobility [5], the discovery of significant places [4] and routines [7] in one's personal life. One factor that has yet been ignored by most work is the crucial role that interpersonal relationships play in shaping one's life. It is known that humans are actively participating in various interactions with others on a daily basis due to their instinct social nature - a comprehensive characterization of individual behaviors for context-aware applications such as social activity prediction and personality inference thus requires an exploration of not only the personal aspects but also the relational aspects captured in humans' social interactions.

A good way to study a user's social interactions is to identify and characterize her *social circles*. This concept is broadly used in online social networks for content organization, e.g, Google+ allows users to explicitly categorize their acquaintances into different circles such as friends, family members, etc. It is natural to think that similar circle categorizations implicitly exist in one's daily interactions in physical world, and are an important type of social context in understanding individual behaviors. As social interactions often happen in close physical distance, a promising type of data that can be leveraged to discover one's social circles in physical world is the time-stamped proximity data continuously sensed by Bluetooth (BT) built in smartphones. Our focus in this note is thus to explore how to automatically identify various social circles of a given user from streams of proximity data sensed by her BT device over long periods of time.

A user in reality often has multiple social circles, each of which consists of a subset of members in her personal social network and is driven by certain unobserved latent aspects. Typical such aspects include "research group", "roommates", "close friends", etc. Many features can be used to characterize an aspect and distinguish it from others, such as interaction time, locations and calling patterns. In this note, we restrict our focus on interaction time and show its power in uncovering and separating meaningful social circles hidden in a user's personal proximity network. Intuitively, a user's interaction with her different circles usually presents distinguishable temporal characteristics, e.g, a user's interaction with her colleagues typically happens on a regular basis during working hours on weekdays, while her stay with close friends can occur more frequently in non-working hours, especially on weekends. We refer to the given user as *ego* and the persons in her BT-sensed dynamic proximity network as *subjects*. Our task then is to discover social circles from a user's ego proximity network, including for each circle its constituent subjects as well as its temporal characteristics reflective of the latent aspect. In review of related work, [6] addressed the similar problem in online social networks based on user profiles, while we adapt it to physical world by leveraging temporal features; [1] mined recurrent group interaction types from

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.

http://dx.doi.org/10.1145/2493432.2493512

BT data, while we adopt an ego-centric perspective on the same data type for better user characterization. Our work is closely related to [3], but has noted advantages, which will be highlighted in the next section.

CHALLENGE AND CONTRIBUTION

We summarize major challenges that motivate our method. In what follows, we abbreviate the *proximity time sequence* of a subject as PTS, which refers to the sequence recording in what time slots that subject is detected as proximate or not to the ego. First, our scenario is purely unsupervised with no circle labels given in the training data. Second, the proximity data sensed by BT is by nature noisy and sparse, since a user may turn off or forget her phone from time to time, and the BT may fail to detect all nearby devices. As a result, a subject may present seemingly peculiar PTS and hence be wrongly classified to an idiosyncratic social circle; multiple subjects who are in the same circle may end up generating notably dissimilar PTS and thus be mistakenly assigned to different circles. Third, neither the latent temporal aspect of each circle nor the circle membership of each subject is available. Either can be trivially gained when the other is known in advance. [3] classifies ego-subject relationship by counting how frequent a subject is proximate to the ego on Saturday night, during conjectured working hours, etc, which implicitly assumes the temporal dimensions underpinning each circle be known *a* priori. This is unreasonable as such temporal dimensions are highly ego-specific. In contrast, we remove this assumption and propose a method to learn both the temporal characteristics and members of each circle simultaneously, which is challenging yet allows for more generality. Fourth, in contrast to [3], we assume one subject may belong to multiple circles of the ego, e.g, it is common that one subject is both the colleague and close friend of a user. Under such an "overlapping circle" condition, one subject's PTS may be explained by a mixture of multiple aspects, the separation of which is challenging, which we will tackle by proposing a novel method that well exploits the fact that each latent circle normally has a single and coherent semantic.

LEARNING APPROACH

To grapple with the aforementioned challenges in our task, we propose a novel unsupervised method motivated by the idea of collaborative filtering. Its spirit is to take a generative approach to explain the proximity-related events observed in the data. In particular, a proximity event in an ego's BT data states which subject s is detected as proximate to the ego in what time slot t. Intuitively, s and t in a proximity record are potentially correlated via a social circle semantics as a latent factor, which in spirit is similar to the idea that users and items are linked via latent interest semantics in recommendation systems. Recall that a social circle is driven by a latent aspect and is characterized by its members and active interaction time. On one hand, a particular aspect (circle semantics) is active at certain time slots (e.g, "family" interaction tends to occur at night and on weekends). On the other hand, each subject potentially belongs to certain (possibly multiple) social circles. From a generative perspective, if a subject s and a time slot t "match" in the latent social circle semantics, then it is highly likely that s is observed proximate to the ego at time t in the data. On the contrary, the event that the ego's BT does not detect some subject s at time t (non-proximate event) is likely due to the "mismatch" of the circle semantics between s and t. The key idea then is to find each subject's social circle membership and each time slot's featured circle semantics such that all proximity-related events (including proximate and non-proximate events) in the ego's BT data as a whole can be best explained from a statistical point of view.

Mathematically, this idea is achieved by probabilistically factorizing a sparse matrix linking subjects with time via proximity events in a low-dimensional space. Suppose there are K circle semantics, we associate with each subject s and each time slot t a K-dimensional real-valued latent vector $\boldsymbol{p}_s$ and $\boldsymbol{q}_t$, with the kth element indicating the extent to which s belongs to circle k and circle k is active at t, respectively. Denote S and T as the set of all subjects and all time slots, and let matrix $\mathbf{P}$, $\mathbf{Q}$ encode all $\boldsymbol{p}_s$, $\boldsymbol{q}_t$ as columns, respectively. We model the ith event in ego u's BT proximity data set $\mathcal{D}_u$ as a Binomial random variable X^i with 1 indicating proximate event and 0 indicating non-proximate event. Let $s(i)$, $t(i)$ denote the subject and time slot associated with the ith event. Intuitively, the probability that a proximate event occurs, $P(X^i = 1)$, depends on the extent to which $s(i)$ and $t(i)$ match in circle semantics, which can naturally be modeled as $\sigma(\boldsymbol{p}_{s(i)}^T \boldsymbol{q}_{t(i)})$, where $\boldsymbol{p}_{s(i)}^T \boldsymbol{q}_{t(i)}$ quantifies the extent of their semantic match and the sigmoid function $\sigma(\cdot)$ "squashes" it to be a legal probability in $[0, 1]$ (similar to the trick in logistic regression). The distribution of X^i is then defined as $P(X^i|\mathbf{P}, \mathbf{Q}) = \sigma(\boldsymbol{p}_{s(i)}^T \boldsymbol{q}_{t(i)})^{X_i} (1-\sigma(\boldsymbol{p}_{s(i)}^T \boldsymbol{q}_{t(i)}))^{1-X^i}$. This reflects the idea that it is the extent of circle semantics match between the corresponding subject and time slot that determines the likelihood of a proximate event. By adopting a maximum likelihood estimation approach, we seek $\mathbf{P}$,$\mathbf{Q}$ to minimize the negative log-likelihood of ego u's BT data set $\mathcal{E}(\mathbf{P}, \mathbf{Q}|\mathcal{D}_u) = -\mathbf{log} \prod_i P(X_i|\mathbf{P}, \mathbf{Q})$ which can be derived as $-\sum_{s\in S, t\in T}(A_{st}\mathbf{log} y_{st} + B_{st}\mathbf{log}(1 - y_{st}))$, where $y_{st} = \sigma(\boldsymbol{p}_s^T \boldsymbol{q}_t)$, $A_{st} = \sum_{i\in R_{st}} X^i$, $B_{st} = \sum_{i\in R_{st}}(1 - X^i)$, $R_{st} = \{i|s(i) = s, t(i) = t\}$. In essence, this approach collaboratively exploits all subjects' proximity data and advocates the sharing of temporal structure information between different subjects. It thus well compensates the sparsity in a single subject's proximity data by using the data of other subjects who potentially share similar circle membership. After learning, the temporal characteristics of social circle k is encoded in the kth row of matrix $\mathbf{Q}$, and the circle membership of subject s can be derived by applying softmax function to $\boldsymbol{p}_s$ such that $P(s \in circle\ k) = e^{\boldsymbol{p}_s(k)} / \sum_{k'} e^{\boldsymbol{p}_s(k')}$. Since $\mathbf{P}$ and $\mathbf{Q}$ are optimized simultaneously in search for a good explanation of the data, we are essentially learning for each circle its members and the temporal dimension along which it is formed simultaneously, thus well addressing the third challenge. In our matrix factorization model, a key point to tackle the fourth challenge (i.e, to allow multi-circle membership of a subject) is to "force" different social circles to form along different temporal dimensions, which makes sense as an ego's interactions with her different social circles seldom overlap in time (e.g, "colleague" and "family" interactions

occur in notably different time). This is achieved by imposing l_1 norm on each $\boldsymbol{q}_t$ to encourage sparse profile. As a result of this sparsity-favored regularization, each row in $\mathbf{Q}$ accounts for a single semantically coherent social circle aspect, which is what we desire. In addition, we add another regularization to encourage neighboring time slots to confer similar circle semantics, based on the intuition that the active time for a social circle interaction is usually piecewise continuous rather than sporadic. The final regularization term is then $\mathcal{R}(\mathbf{Q}) = \sum_{t\in T} ||\boldsymbol{q}_t||_1 + \frac{1}{2}\sum_{t\in T} ||\boldsymbol{q}_t - \boldsymbol{q}_{t-1}||_2^2$. The final optimization problem is thus to minimize the objective function $\mathcal{E}(\mathbf{P}, \mathbf{Q}|\mathcal{D}_u) + \lambda\mathcal{R}(\mathbf{Q})$ w.r.t $\mathbf{P}$ and $\mathbf{Q}$, where λ controls the regularization strength. The optimization is conducted using gradient descent algorithm, which alternates between updating $\mathbf{P}$ and $\mathbf{Q}$ until convergence. The relevant gradients are derived as follows by applying chain rules, where we have used the fact that $\sigma'(a) = \sigma(a)(1-\sigma(a))$.

$$\nabla_{\boldsymbol{p}_s}\mathcal{E} = -\sum_{t\in T} \boldsymbol{q}_t(A_{st} - (A_{st}+B_{st})\sigma(\boldsymbol{p}_s^T\boldsymbol{q}_t)) \quad (1)$$

$$\nabla_{\boldsymbol{q}_t}\mathcal{E} = -\sum_{s\in S} \boldsymbol{p}_s(A_{st} - (A_{st}+B_{st})\sigma(\boldsymbol{p}_s^T\boldsymbol{q}_t)) \quad (2)$$

$$\nabla_{\boldsymbol{q}_t}\mathcal{R} = [sign(\boldsymbol{q}_{t,k})]_{k=1\ldots K} + 2\boldsymbol{q}_t - \boldsymbol{q}_{t+1} - \boldsymbol{q}_{t-1} \quad (3)$$

The algorithm can easily be adapted to the online setting. When new proximity time data comes in, the sufficient statistics A_{st} and B_{st} are incrementally updated, which will then replace the old ones in (1) and (2) to continue updating the parameters from their current values using gradients. For complexity, our algorithm is linear in terms of the number of subjects and time slots, and thus can scale to large data sets.

EXPERIMENTS

For repeatability, we carry out experiments on the public Reality Mining dataset [2] on 95 academic users over 9 months. The BT in these users' Nokia mobile phones periodically scanned and recorded nearby devices of their MAC addresses and the time at 5-minute interval, which generated for each user a large set of proximity events that can well fit into our learning task. Notably, a comprehensive survey on the studied users was provided, which reveals quite useful personal and relational information that can be used for verification.

Qualitatively, we first show what ego-centric social circles are learned and explain by results how our model overcomes the mentioned challenges. We encode the combination of each day of week and each hour of day as a time slot. For an ego, the temporal profile of each of her circles can be read from the rows of matrix $\mathbf{Q}$. For illustration, we select user 6 as the ego (a graduate student from MIT Media Lab), set K to 6, apply our model, and plot 4 rows in his matrix $\mathbf{Q}$ learned from his BT proximity data using heatmap in Figure 1. Each row is reshaped as a matrix with x axis encoding the time of day and y axis encoding the day of week. The bright areas in each matrix indicate the dominating [1] active time of each circle. Figure 1 clearly shows 4 typical interpretable social circles in user 6's daily life. Figure 1(a) apparently indicates the circle of user 6's lab colleagues, as the proximity time mostly falls on weekdays during working hours. Interestingly, this circle's active time also reflects user 6's personal working schedule (roughly from 11am to 8pm) which agrees with the survey. In Figure 1(b), the dominating weights are restricted in the morning before working hours on weekdays and cover the whole afternoon on weekends, which indicates the circle of close friends with whom user 6 stays mostly in his leisure time. Combining (a) and (b), it is interesting to note that on weekdays user 6 "switches" between his two circles around his start working time (which might have daily variation), and that he doesn't meet friends until afternoon on weekends. This demonstrates that even in the absence of personal data such as traces, by merely monitoring a user's proximity relationships with others, we are still able to reveal her personal routine behavior pattern. Such a routine becomes more complete for user 6 when his third circle in Figure 1(c) is considered. Roughly complementary to circle 1, this circle is clearly about family member, or roommate interactions typically happening at nights and sometimes in the daytime on weekends. Figure 1(d) indicates a weekly event, probably a group meeting in which user 6 interacts with his research group on a weekly basis. This example shows that besides personal data, mining a user's social interactions can also achieve clear individual behavior characterization, and maybe in a more fine-grained manner. Note that our method is superior to [3] in that the active time of each circle is learned from data rather than assumed as prior knowledge. We next turn our focus to the social circle membership of the subjects to

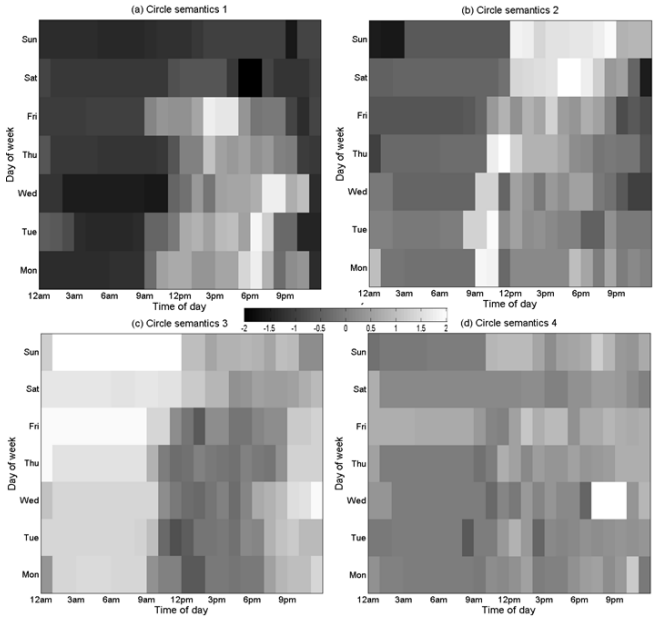

Figure 1. Temporal profiles of 4 typical social circles of ego 6

[1]Unless $\mathbf{Q}$ is constrained to be positive (making the optimization problem more difficult), dominating factors in $\mathbf{P}$ and $\mathbf{Q}$ can be either very positive or very negative due to model symmetry, making the learned results more difficult to interpret in theory. In practice, we circumvent this problem by initializing $\mathbf{P}$ and $\mathbf{Q}$ as random small positive values, such that the gradient update will tend to increase both $\boldsymbol{p}_s$ and $\boldsymbol{q}_t$ initially for dominating factors (circle memberships and active times), and hence will produce results where large positive values indicate dominating factors, as shown in Figures 1&2.

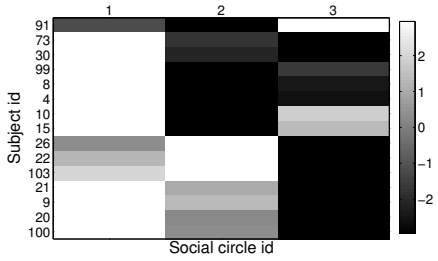

Figure 2. Social circle membership for selected subjects

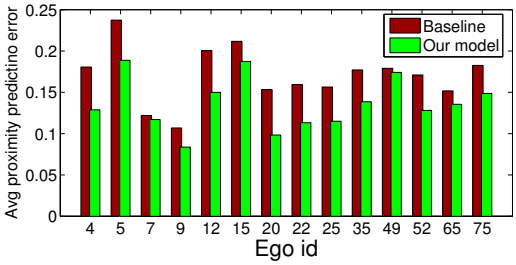

Figure 3. Subject proximity prediction performance for selected egos

whom user 6 is socially related. Such knowledge is encoded in the latent vector $\boldsymbol{p}_s$ of each subject. The indices of the dominating values in $\boldsymbol{p}_s$ indicate the most probable circles to which subject s belongs. We select some subjects and plot in Figure 2 parts of their circle membership vectors (all normalized to the same range) w.r.t the first 3 circles, from which we can see which subjects each circle is comprising of. A key innovation in our method is its ability of discovering overlaps between circles, e.g, the first two dominating factors in $\boldsymbol{p}_{103}$ implies that subject 103 is both the colleague and close friend of user 6. Subject 15 and 10 belong to both circle 1 and 3 with high likelihood, suggesting that they are both working in the same group with user 6 and living close to him, a fact that can be verified by checking the raw PTS of subject 15 or 10 with ego 6 to find that they stay with 6 both during working hours and in late night. However, the mere PTS of a subject does not allow for such a semantically reasonable characterization as it might mingle multiple circles' active time and cannot be separated when the temporal dimensions along which each circle is formed are unknown. Owning to the collaborative idea and the usage of sparsity-favored regularization on $\boldsymbol{q}_t$, our method is able to learn for each circle a coherent semantic (e.g, "colleague" semantics is more desirable than a semantics mixing "colleague" and "friends") and naturally explain a subject's proximity time sequence as a mixture of them, thus enabling more interpretable user characterization.

Quantitatively, we evaluate our model's ability in predicting which subjects are likely to be proximate to an ego user at a given time, and compare it with a baseline. In particular, we split a given ego's BT data into training set and test set, and make use of the knowledge (i.e, $\mathbf{P}$ and $\mathbf{Q}$) learned from training set to predict for each time slot the probability that each subject is proximate to that particular ego. The predicted result of a time slot is then compared with the "groundtruth" in the test set (by frequency counting) to measure the prediction error for that particular time slot. The average of the prediction errors over all time slots is then used to measure the model performance for that particular ego. The baseline we choose predicts the probability that a subject is proximate at a given time slot by counting the frequency of such events in training set using only that particular subject's proximity data. The key difference between our model and the baseline is that in our model, the prediction of the likelihood that a subject is proximate at a given time is based on whether the subject and the given time match in social circle semantics, a knowledge that is learned by exploiting all subjects' historical data in a collaborative way, while the baseline only uses the target subject's historical data to make that prediction. We select certain users as egos and compare our model with the baseline in their subject proximity prediction performance in Figure 3. It can be seen that our method gives superior results, due to its awareness of the underlying social circle semantics and its collaborative nature. But the key advantage of our method over the baseline is that our method can uncover and represent circle semantics along subject and time dimensions (Figures 1&2) while the baseline cannot.

CONCLUSION AND FUTURE WORK

In this note we proposed a learning framework to discover ego-centric social circles from BT-sensed proximity data, based on the intuition that it is the unobserved social circle semantics that decides which subjects are likely to be proximate at when. Experiments showed how the social circle semantics uncovered can be leveraged for personal life interpretation and social event prediction. We believe that modeling mobile social context is becoming increasingly important, and the fusion of it with personal mobile data analysis will prove promising in building advanced human-centered ubiquitous systems. In the future, we plan to extend the work to other more practically accessible location data like Foursquare check-in data, where new challenges arise as the proximity relationships are not directly available but should be inferred from check-in logs. Another extension is to model how a user's social circles evolve over time, which can offer crucial insights into users' social behavior dynamics.

ACKNOWLEDGEMENT

We thank Huawei Corp. Contract YBCB2009041-27.

REFERENCES

1. T. M. T. Do and D. Gatica-Perez. Human interaction discovery in smartphone proximity networks. *Personal and Ubiquitous Computing*, pages 1–19.
2. N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.
3. N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.
4. K. Laasonen, M. Raento, and H. Toivonen. Adaptive on-device location recognition *Pervasive Computing*, pages 287–304, 2004.
5. M. Lin, W.-J. Hsu, and Z. Q. Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proc. UbiComp 2012*, pages 381–390. ACM, 2012.
6. J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Proc. NIPS 2012*, pages 548–556, 2012.
7. J. Zheng and L. M. Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proc. UbiComp 2012*, pages 153–162. ACM, 2012.