Socially Relevant Venue Clustering from Check-in Data

Yoon-Sik Cho
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
yoonsik@isi.edu

Greg Ver Steeg
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
gregv@isi.edu

Aram Galstyan
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
galstyan@isi.edu

ABSTRACT

The recent proliferation of location-based social network services has resulted in an abundance of spatial-temporal data on user mobility. Understanding individual and collective mobility patterns is important for many applications. In this study, we examine the similarity of users based on the venues they have visited in the past. In contrast to the previous approaches that measure user similarity based on co-location patterns, here we first cluster venues in some latent (lower-dimensional) space, which allows us to capture the similarity between two users who have not necessarily visited the exact same venues in the past. We validate our approach on real-world data and demonstrate an improved performance over previous methods.

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory**]: [Information theory]; H.2.8 [**Database Applications**]: [Data mining, Spatial databases and GIS]

Keywords

Location Based Social Networks, Clustering, Information Bottleneck

1. INTRODUCTION

Despite the efforts of social scientists, understanding human mobility patterns remains a challenging problem. As sensors become more ubiquitous with accelerometers and GPS embedded in cell phones, computer scientists are able to analyze movements on a more fine-grained level. The emergence of location based social network services take the potential even further. Large scale data covering wide areas over long timescales is coupled with detailed information about users' online interactions.

Location Based Social Networks (LBSN) bear unique features. Many of the application interfaces now enable users to select the location they would like to check into from an automatically generated list. Often the lists are displayed by tracking the GPS coordinates of a current user and searching for nearby candidates. If the name of the location is not provided by the service, the user can add it for future visitors. This feature is remarkable in this area of studies in that the exact location can be pin pointed using the labels attached to each of the check-ins. The noise of GPS coordinates can be filtered easily using these labels. This enables us to collect the users who have visited specific venues during a given time period.

Another unique feature in LBSN is the sharing of user locations with friends. On popular social networking sites like Facebook, users can pinpoint where their friends are and where they have been in the past if they have checked-in. This creates on-line influence which is the social network equivalent of word-of-mouth influence. Each check-in creates a visible reminder that may induce friends to return to a location or visit for the first time. In this sense, network structure may have major implications in this area of study.

Understanding and modeling individual movement has many applications. By understanding each movement and the relationship between users, service providers could recommend some venues to a group of users with potential interest. In a broader context, human mobility models also impact prediction of the spread of disease, controlling traffic congestion, business marketing, and urban analysis.

Yet another huge impact on other fields of study comes from the fact that the venues comprise not only the geo-coordinate information but also other useful information when combined with the profile of users who have visited it. This may provide information about the behavior patterns associated with a venue or the groups which frequently visit the venue. For instance, based on the location, timing, and composition of a group, it could be possible to infer the activity as 'studying' with high probability. This intuition suggests that venues see similar patterns of activity based on the users who visit.

In this work, we use the network structure information to cluster venues so that a venue's group reflects its functionality. This coarse representation of venues may be useful in many ways but we focus on two concrete benefits. First, through clustering the venues, we may be able to have a better understanding of what venues represent as a whole. For instance, the venues connected to schools, libraries, bookstores could be may be related to activities like 'studying'. Second, with coarse representation of venues the unknown relationships among users can be inferred based on the set of venues two users have visited. Using LBSN dataset, we

Eleventh Workshop on Mining and Learning with Graphs. Chicago, Illinois, USA

found that many of the actual friends show similar check-in patterns, but some of the pairs had no overlap at all. We show that through clustering venue types our model can correctly infer relationships even between pairs of users with no overlap of venues.

We present network infused agglomerative information bottleneck, which is an extension of agglomerative information bottleneck [11]. This simple non-parametric method allows us to cluster venues utilizing the network information. As mentioned previously, we show two advantages: categorizing venues and edge prediction using these clustered venues. However, for ease of validation, we mainly focus on edge prediction in this work and show how coarse representation performs better than using the raw venue data. We also show how our clustering performs better relative to other methods of dimensionality reduction.

2. RELATED WORK

There have been a number of studies [2, 4, 8, 13] using geo-spatial dataset for modeling mobility patterns in social networks. For instance, Ref. [13] defined *mobile homophily* based on visitation frequencies, and used this measure to infer social interactions. The difference between our approach and the prior work relies on measuring the homophily. Namely, our method first projects venues onto latent space, and then finds similar users in this space. Thus, our approach can yield high similarity for a pair of users who have never visited the same venue, provided that they have visited similar venues. This is very different from approaches that use, for instance, distances between locations for link prediction.

As we mentioned above, we intent to compress the check-in data into a coarser representation. There are many existing approaches for dimensionality reduction. Latent Semantic Analysis (LSA) [7] and Latent Dirichlet Allocation (LDA) [1] were originally introduced in NLP to discover hidden concepts/topics characterizing document data using the term-document matrix. Probabilistic Matrix Factorization (PMF) [9, 10] that was developed for collaborative filtering works by decomposes the rating matrix two matrices and works well on predicting missing rates. Spectral Co-Clustering [5] simultaneously clusters row and column using spectral graph partitioning.

We would like to note that, similar to our work, Ref. [6] used LDA to cluster venues, although they did not use the obtained clusters for link prediction. Furthermore, in contrast to [6] (and the dimensionality reduction methods listed above) the approach proposed here is a supervised method as it uses information about known social ties.

3. DATA DESCRIPTION

We use Gowalla dataset [2] in this work. Gowalla is a location based social network (LBSN) service where each user can post their current location and share it with its friends. The 'check-in' consists of the node id(user), the actual date and time, and coordinate with the location id provided by Gowalla. Location ID becomes useful when two locations which are close or have the same coordinates but located on different building levels need to be distinguished. To distinguish from previous works using geo-space dataset, we use *venues* which is a location with ID in LBSN, instead of *locations*, which were mostly based on cell towers using cell phone data.

The majority of the users in the dataset show sparse activity in 'check-in's, which makes the modeling difficult. We observe that 20% of the most-active users were responsible for 80% of all check-ins. For the experiment, we use the check-in data of users from three major cities in USA, which includes San Francisco, Austin, and New York. For each city, the check-in data is encoded as an $|V| \times |U|$ matrix, where V is the set of venues and U is the set of users. The (i, j)-th entry indicates the number of visits of user j to the venue i. The dataset also contains node-node friendship matrix, which is an unweighted and undirected graph. We use the check-ins of users who showed active histories in those three cities.

Measuring Similarity.

In previous work [13], one of the measure used to compute mobile similarity was cosine similarity between two vectors of the given users. Each user had occurrence vector, where the i-th component of the vector counts the number of the visits to location i of the given user. Cosine similarity is an inner product of two l_2-norm vectors, which measures the cosine of angle between them:

$$\mathrm{SIM}_{\cos}(A, B) = \frac{A}{\|A\|} \cdot \frac{B}{\|B\|} \tag{1}$$

Another measure of similarity is Kullback-Leibler (KL) divergence (defined below). Our results indicate that KL divergence achieves better inference on friendships than cosine similarity [1]. Hence we use KL divergence as a metric of measuring similarity of check-in histories.

4. NETWORK-INFUSED AGGLOMERATIVE INFORMATION BOTTLENECK

Our objective in this work is to capture the unknown edges using the similarity between the users in some latent space. Our approach is based on a variation of the Information Bottleneck (IB) method [12]. This method ttys to find a compressed representation $\tilde{X}$ of the original data X so that $\tilde{X}$ still contains useful information about some relevance variable Y. In other words, IB tries to find the features of the original dataset that are most useful for predicting the relevance variable Y, while discarding the features (via compression) that are not.

To make this intuition more formal, let us recall the definition of the mutual information between two random variable X and Y:

$$\begin{aligned} I(X;Y) &= \sum_{x\in X, y\in Y} p(x,y)\log\big(\frac{p(x,y)}{p(x)p(y)}\big) && \text{(2a)}\\ &= \sum_{y\in Y} p(y)\mathrm{D}_{\mathrm{KL}}(p(x|y)||p(x)) && \text{(2b)}\end{aligned}$$

where in the second equation we have introduced the Kullback-Leibler (KL) divergence:

$$\mathrm{D}_{\mathrm{KL}}(p||q) = \sum_i p(i)\log\big(\frac{p(i)}{q(i)}\big) \tag{3}$$

[1] We believe this is mainly due to the l_1 normalization. l_2 norm is sensitive to the scaling factor especially when dealing with high dimensional vectors, where as l_1 vectors shows more robustness [3].

The objective of IB method is to find a compact representation $\tilde{X}$ of the original variable X that results in a minimal loss of information about the relevance variable Y. Introducing a Lagrange multiplier β, the above objective is captured by the following functional:

$$\mathcal{L}[p(\tilde{x}|x)] = I(X;\tilde{X}) - \beta I(\tilde{X};Y) \tag{4}$$

Note that IB can be viewed as a soft clustering algorithm characterized by the conditional distribution $p(\tilde{x}|x)$. When the compressed representation has finite cardinality, then in the limit $\beta \to \infty$, IB reduces to hard clustering [11]. In this limit, the first term in Equation 4 is discarded, and the problem reduces to maximizing $I(\tilde{X};Y)$. We focus on this scenario from now on.

We adapt a version of IB known as agglomerative Information Bottleneck [11], which is essentially a bottom-up hard clustering method. The agglomerative IB starts with trivial partition where each data point is in its own cluster. Define *information loss* as the decrease in the mutual information $I(\tilde{X};Y)$ due to merging, $\delta I_y = I(X;Y) - I(\tilde{X};Y)$. Then, at each iteration, the agglomerative IB greedily merges two clusters that have the minimum mutual information loss.

We now define an objective function inspired by the IB approach. In our case, the data that we would like to compress is the set of all venues, while the relevance variable is the existing network structure. Then, intuitively, we would like to compress the venues so that the users who are linked in the network will be *close* to each other in the compressed representation, whereas the users who are not linked in the network will be further away. Thus, we define

$$I_S(X;Y) = \sum_{y\in Y} p_w(y) \Big\{ \mathrm{D}_{\mathrm{KL}}(p(x|y)||p_{\mathcal{S}_y}(x)) - \mathrm{D}_{\mathrm{KL}}(p(x|y)||p_{\tilde{\mathcal{S}}_y}(x)) \Big\} \tag{5}$$

where $\mathcal{S}_y$ denotes the set of friends of user y, and $\tilde{\mathcal{S}}_y$ denotes the set of non-friends of user y, and $p(x|y)$ is the probability of a given user visiting venue x.

The two terms in the objective functions are the results of combining two types of information - existence and absence of links between the users. Since our objective is to differentiate the two sets (friend set vs non-friend set) for each users, we separate the two by penalizing the other term, the distance to the probability of non-friends. Note also that instead of $p(y)$, our objective uses $p_w(y) = \frac{\#\text{of edges containing } y}{\#\text{of edges}}$ which gives more weight to the users that have more edges.

Having defined the above objective, we can use the greedy bottom-up technique defined above to find hard clustering of the venues. The following remark is due: Since users do not visit all the venues, the denominator in the KL divergence is often zero. To avoid this, below we use Jensen Shannon (JS) divergence, a symmeterized and smoothed version of KL divergence:

$$J(\mathcal{C}) = \sum_{u\in U} w_u \{ JS(\mathbf{p}_u^{\mathcal{C}}||\mathbf{p}_{\mathcal{S}_u}^{\mathcal{C}}) - JS(\mathbf{p}_u^{\mathcal{C}}||\mathbf{p}_{\tilde{\mathcal{S}}_u}^{\mathcal{C}}) \} \tag{6}$$

$$JS(\mathbf{p}||\mathbf{q}) = \mathrm{D}_{\mathrm{KL}}(\mathbf{p}||\mathbf{r}) + \mathrm{D}_{\mathrm{KL}}(\mathbf{q}||\mathbf{r}), \text{ where } \mathbf{r} = \frac{1}{2}\mathbf{p} + \frac{1}{2}\mathbf{q} \tag{7}$$

$$\mathbf{p}_u^{\mathcal{C}}(k) = \frac{\#\text{of user } u \text{ visiting cluster } k}{\#\text{of total check-ins of user } u} \tag{8}$$

Algorithm 1 Venue Clustering

```
**Size:** consider total of $|V|$ venues, $|U|$ users
**Input:** $|V|$ by $|U|$ co-occurrence matrix $\mathbf{X}$, and
$|U|$ by $|U|$ adjacency matrix $\mathbf{Y}$, which is partially observable
**Initialization:** Start with $|V|$ clusters, each of which contains a venue
**repeat**
  **for** $i, j = 1$ **to** $C$, i<j **do**
    $d_{ij} = J(\mathcal{C}) - J(\bar{\mathcal{C}})$
    where $\bar{\mathcal{C}} = \big\{\mathcal{C} - \{c_i, c_j\}\big\} \cup \bar{c}_{ij}$,
    and $\bar{c}_{ij}$ is a merge of cluster $i$ and $j$
  **end for**
  **Merge:**
  – Find $\{\alpha, \beta\} = \arg\min_{i,j} d_{ij}$
  – Merge $\{c_\alpha, c_\beta\} \to \bar{c}_{ij}$
**until** $\min d_{ij} < -\epsilon$
**Output:** $C$ by $U$ matrix $\mathbf{X}_C$
```

In Equation 6 and 8, $\mathcal{C}$ represents the set of the current clusters; The component of the probability vector is defined in equation 8. The overall procedure is shown in Algorithm 1.

5. EXPERIMENTAL RESULTS

For our experiments we focused on Gowalla data from three major US cities - San Francisco, Austin, and New York. Those three cities exhibited more active 'check-ins' compared to other major cities. For each city, we used top 20% active users of which the check-ins form 80% of all check-ins. We also left out the venues with fewer than 10 different visitors during the considered period.

5.1 Venue clustering

In our first set of experiments, we examined whether our approach yields meaningful clustering of the venues. We run our algorithm starting with the clusters where each venue constitutes a separate cluster initially. We merge those clusters following the algorithm 1, until the information loss is less than a predefined threshold.

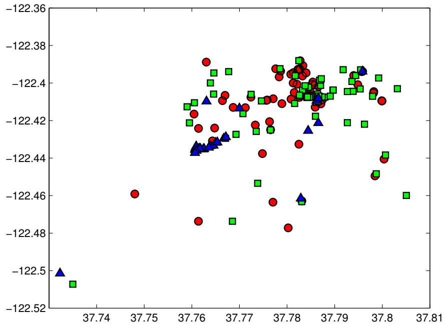
For all three cities, we examine the top three clusters that contain the most venues. Since we are only interested in what the clusters represents, we use 100% of the edge information between the users for the clustering. For each cluster we find the top 10 popular venues. We assume that the number of unique users in each venue represents the venue popularity, i.e., the more unique visitors the more popular it is. Top 10 venues of each clusters with most unique users are examined.

The name of the venues in top three clusters of San Francisco are presented in Table 1. We observe that the cluster C_1 mainly consists of amusement facilities such as theater, brewery (or bar), and cafe. Cluster C_2 mostly has shops and venues in the shopping district of San Francisco. And for the cluster C_3 we found that most of the venues seem to be associated with the LGBT (lesbian, gay, bisexual, and transgender) community. Thus, we see our clustering algorithm is able to capture semantic information about venues using the friendship network information between the users who check-in.

Figure 1 shows the actual mapping of the venues. It is seen that the venues that belong to the same cluster are not

Table 1: Top 3 largest clusters in San Francisco

C_1	Metreon(movie,video game, city target)
	Amendment Brewery
	Mint Plaza
	SFMOMA
	Moscone Center North
	San Francisco Ferry Bldg
	Whole Food Market
	Thirsty Bear Brewery
	Sightglass(coffee)
	Bloodhound (bar)
C_2	Apple store
	Union Square Park
	Westfield San Francisco Centre(shopping mall)
	Moscone West (exhibition hall)
	Powell st (cable car turntable)
	Flood Building (powell st shopping district)
	Transamerica pyramid
	(small shop)
	Lucca Delicatessen(italian deli brocery)
	Macy's
C_3	Rainbow World Fund(friends community)
	Toad Hall (bar)
	Building (rainbow flag)
	Sunflower Cafe
	(unknown, moved out)
	QBar (rainbow flag)
	Moby Dick (bar, rainbow flag)
	440 Castro (underwear night, rainbow flag)
	closed property (rainbow flag)
	Exclusive club



(a)

Figure 1: Geo plot of three clusters C_1(red), C_2(green), C_3(blue)

necessarily geographically close to each other. Instead, the closeness is in the induced latent space. We also observe that C_3 is more localized geographically compared to other clusters. This is because many of the venues in this cluster are located in a prominent LGBT neighborhood in the city.

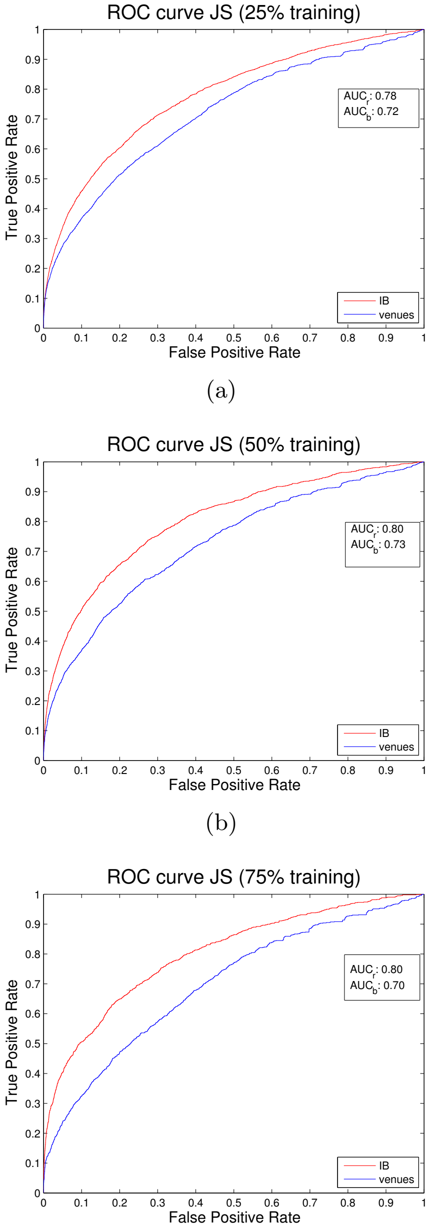
5.2 Reconstructing Edges

In the next experiment we use the coarse-grained representation of the venues to predict social links among the users. In the San Francisco dat set, there were $3,360$ edges out of $706,266$ pairs. Furthermore, out of $1,680$ edges, 565 (more than third) had no common venues between the two at all. The New York dataset has $1,205$ active users with $1,051$ venues, with $1,781$ edges from $725,410$ pairs. And for the Austin datset, there were $1,920$ active users with $9,126$ edges.

For the experiment, we use fraction of the existing edges to cluster the venues using our algorithm. We then try to recover the remaining edges based on the venue clusters they have visited.

As our algorithm uses available network information in venue clustering, we expect to achieve better accuracy with more network information. To examine this effect, we control the observable edge ratio to $25\%, 50\%$, and 75% and infer the $75\%, 50\%$, and 25% of edges respectively. Due to the limited space, we only present the results from the San Francisco data (the results were similar for the remaining two datasets). In Figure 2 we show the ROC curve for different approaches, together with the corresponding AUC scores. We see that measuring user similarity based on clusters of venues results in more accurate link prediction. In other words, the induced latent categories of the venues are a better measure of similarity than the individual venues themselves.

We also compare our algorithm to other baselines described in Section 2. All of the baselines uses reduced dimensional representations of venues for finding similarities between users. For the experiment, we used 50% of the edge information for our algorithm, and validated with the other 50% of the data as a test set assuming the edge information is unknown. With the same test set, we inferred the edges between users using other baseline methods and compared it to ours. As shown in Figure 3, our method (IB) outperforms other baselines. (topic: LDA, co-clustering: Spectral



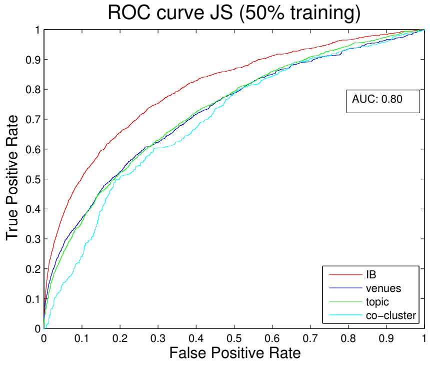
(a)

(b)

(c)

Figure 2: ROC curve for link prediction using JS-divergences. The results are shown for the information bottleneck (red) and unclustered original venues (blue) on varying size of training set (25% of whole data, 50% and 75%). We also show the corresponding AUC scores.

co-clustering). We note, however, that direct comparison of the methods is a little unfair, since our method makes use of additional (social network) information for clustering the venues, whereas the baselines above are fully unsupervised. Nevertheless, our results clearly indicate that information about social interactions is indeed relevant for clustering venues.



(a)

Figure 3: ROC curve using JS-divergence compared to other baselines. The AUC is for the red plot (IB) only. IB denotes our model, where as the venue denotes the edge reconstruction using the unclustered venues. topic: LDA and co-cluster: Spectral co clustering are the baseline we have introduced previously

6. CONCLUSIONS

Finding similar users or friends in LBSN is important for better understanding user mobility patterns. Though there are many venues in a city, only a small number of venues are visited by individual users. The mobility patterns (i.e., the venues that a user frequently checks-in) exhibit the characteristics of each user. Conversely, we can predict the venues that the users might be interested based on our inference. Reaffirming many previous studies, the social network plays a great role in inferring the characteristics of users. In this work, we focused on reconstructing the social network based on other observable network and the check-ins of users. We showed that when the venues are merged using our algorithm, we achieve better predictions about the social network. We also validated that our cluster contains meaningful representation by examining the name of the venues and the actual locations on geo-space.

Acknowledgments

This research was supported in part by the US AFOSR MURI grant FA9550-10-1-0569, US DTRA grant HDTRA1-10-1-0086, and DARPA grant No. W911NF-12-1-0034.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD'11*, 2011.

[3] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley-Interscience, New York, NY, USA, 1991.

[4] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, Dec. 2010.

[5] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM.

[6] K. Joseph, C. H. Tan, and K. M. Carley. Beyond "local", "categories" and "friends": clustering foursquare users with latent "topics". In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 919–926, New York, NY, USA, 2012. ACM.

[7] T. K. Landauer and S. T. Dutnais. A solution to platoÕs problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240, 1997.

[8] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *WSDM'12*, 2012.

[9] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the International Conference on Machine Learning*, volume 25, 2008.

[10] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

[11] N. Slonim and N. Tishby. Agglomerative information bottleneck. In *NIPS*, pages 617–623. MIT Press, 1999.

[12] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[13] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.