Location^3: How Users Share and Respond to Location-Based Data on Social Networking Sites

Jonathan Chang and Eric Sun
1601 S California Ave.
Palo Alto, CA 94304
{jonchang, esun}@fb.com

Abstract

In August 2010 Facebook launched *Places*, a location-based service that allows users to check into points of interest and share their physical whereabouts with friends. The friends who see these events in their News Feed can then respond to these check-ins by liking or commenting on them. These data consisting of the places people go and how their friends react to them are a rich, novel dataset. In this paper we first analyze this dataset to understand the factors that influence where users check in, including previous check-ins, similarity to other places, where their friends check in, time of day, and demographics. We show how these factors can be used to build a predictive model of where users will check in next. Then we analyze how users respond to their friends' check-ins and which factors contribute to users liking or commenting on them. We show how this can be used to improve the ranking of check-in stories, ensuring that users see only the most relevant updates from their friends and ensuring that businesses derive maximum value from check-ins at their establishments. Finally, we construct a model to predict friendship based on check-in count and show that co-check-ins has a statistically significant effect on friendship.

Introduction

The recent rise in popularity of location-aware mobile phones has brought about a corresponding emergence of applications that allow users to check-in to points of interests (POIs) and thereby share their locations with their friends. These services have enabled the collection of a new kind of data about Internet users, facilitating many new products and analyses that were previously not possible without such precise, temporal data.

Past research has mainly focused on analyzing location data collected from either indirect sources of location or via small-scale experiments on users. (Crandall et al. 2010) analyzes a dataset of 38 million geo-tagged photos from Flickr and finds that two people have a 60% chance of being friends when they have five co-occurrences within a day in distinct cells with sides equal to 1 latitude-longitude degree. The authors point out and attempt to control for several potential sources of bias in this data: for example, people tend to take photographs when out with friends, thus increasing the proportion of social ties among observed co-occurrences; users may seek contacts on Flickr by explicitly searching for people who have geo-temporally co-occurred with them; and photos must be geo-tagged in order to be included in their analysis. On the other hand, past work on explicitly-provided location data has been limited as well. (Tang et al. 2010) breaks location sharing into categories of purpose-driven sharing vs. social-driven sharing and examines the effect on users' decisions on privacy. (Lin et al. 2010) uses extensive data collected from a small set of individuals to construct a machine-learned model of how users refer to places by name. This can be important because no location-based service has a perfect database of POIs and thus must rely on users to help create places that are missing from the provider's database. Further, there has also been work bridging the geographical/social gap using graphical models of social networks (Bonato, Janssen, and Pralat 2010; Scellato et al. 2010). However, because of the novelty of the field and lack of broad data availability, there has not yet been a large-scale analysis of explicitly-provided location data using a service's entire set of check-ins.

This paper analyzes a dataset of check-in and POI data collected from *Places*, Facebook's mobile check-in product. We construct a model to predict where people will go in the future given various demographic data as well as data about their surrounding places. Then we predict how users' friends will respond to these check-ins. Finally, we develop a model to determine whether friendship can be inferred from check-in data.

With over 500 million active users, including over 250 million active mobile users, Facebook is a natural place for people to share data with friends. With the launch of Places, these these data now include fine-grained, temporal data about users' physical locations. Since Places was launched in the United States in August 2010, adoption has increased steadily, and as of the time of this writing over 2 million check-in events are created every day.

When a user arrives at a particular destination, she may choose to log her location through Places. This action will be posted to friends' News Feeds, where they may like and/or comment on the action. Furthermore, if she is accompanied by friends, she may tag them in the same check-in story. Through News Feed, the aggregation of friends who have

Figure 1: Heatmap of check-ins in San Francisco, California. Brighter shading denotes areas of higher check-in activity.

checked in to the same location into a single story provides organized information about friends' whereabouts and activities. In addition, Places allows friends who are in close physical proximity to make unexpected and serendipitous in-person connections.

Although only a small percentage of Facebook users are active users of Places, the global reach and popularity of Facebook means that there are sufficient data to make powerful, general predictions of where people will go, how their friends will respond, and how people relate to one another. The results of this analysis may help businesses target deals and advertisements to users of location-based services such as Facebook Places that frequent their establishments as well as their friends. Furthermore, the unique time component of check-in data has powerful implications for mobile advertising: in addition to targeting demographics, the predictions suggested by this analysis also target time and create small and valuable temporal windows for potential customers. For example, advertisements may be targeted to be redeemed in the next few hours as opposed to being available for the next week. As a result, the benefit to users is that advertising can truly be catered to their needs. This benefit extends beyond businesses to local activities and special events. The timeliness of a particular advertisement based on where users and their friends may be represents a new level of relevance of information.

Data

In order to do a deep examination of location interactions, we consider all points of interest in the Facebook Places database within the boundaries of San Francisco, California as of January 31, 2011. In the analyses that follow, all data were analyzed in aggregate, and no personally-identifiable information was used. The dataset was further split into a training set ending on January 24, 2011; the balance was used as a test set. Figure 1 is a heatmap depicting the check-ins in our dataset. Brighter areas represent regions with a higher incidence of check-ins.

Predicting Where People Will Go

In this section, we address the problem of predicting where a user will check into next, given information about users, nearby places, and people who have checked into these places. Predicting where a user will check into next can drive applications such how to rank places on users' mobile devices, to suggest new places for users to check into, or to provide businesses with ad-targeting data. Additionally, it yields insights into how users check into places and complements traditional places analytics.

We describe the features used to make predictions in the first subsection along with their descriptive statistics. We then present the model and its performance on the dataset.

Features

Our model generates the probability of checking into a place give the following features:

Previous check-ins The number of times the user has previously checked into the place, along with the total number of times the place has been checked into by all users.

Friend check-ins The number of times any of the user's friends has checked into the place.

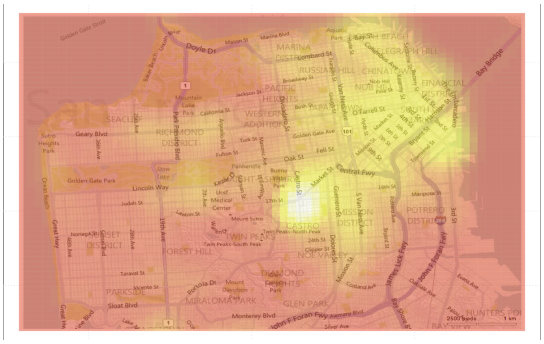
Demographic data The number of times people of the user's age and gender have checked into the place. For example, Figure 2 shows heatmaps of where males and females are especially likely to check in (that is, where the check-in probability of that gender divided by the check-in probability of both genders is large). Places in the Castro neighborhood have an unusually high number of check-ins from males, while females tend to check into the Union Square neighborhood.

Time of Day The number of times users have checked into each place at each hour of the day and day of the week. Figure 3 shows heatmaps of where users of the product are likely to check in in the morning (6 A.M. to 10 A.M.) versus the evening (8 P.M. to midnight).

User membership in a low-dimensional representation of the place space Latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) is a mixed-membership model which has been successfully used to detect patterns and clusters in text. It presumes data that consist of a set of documents, each of which contains an unordered set of words, and finds a set of *topics*, sets of words which tend to co-occur. It also assigns to each document a vector representing how much each document participates in each topic. This model is depicted in Figure 4.

For our application we consider each user a "document" and the set of places a user has checked into as its "words." Thus LDA gives us groupings of places which tend to occur (topics), and assigns to each user a vector indicating how often they check into places in each topic. We use this latter vector as a feature for our model.

Note that unlike the clustering used in (Cranshaw and Yano 2010) to find neighborhoods, our approach here uses



(a) male

(b) female

Figure 2: Heatmaps of check-ins broken down by gender. Brighter shading denotes areas of higher check-in activity.

(a) 6 A.M. to 10 A.M.

(b) 8 P.M. to midnight

Figure 3: Heatmaps of check-ins broken down by time of day. Brighter shading denotes areas of higher check-in activity.

the co-check-in as the unit of clustering rather than co-location. However, the clusters found in our approach often do correspond to conceptual neighborhoods. Table 1 shows a few of the topics inferred by the approach. The first two topics correspond clearly to neighborhoods (the Castro and the Marina). The last contains geographically dispersed places (the Richmond, Sunset, and Chinatown neighborhoods) which all have high Asian-American populations. Thus, this approach has the potential to find culturally similar "neighborhoods" even if they span different areas of a city.

Distance of place to user's usual location The distance between the place and a proprietary determination of the user's usual location whence he or she uses Facebook.

Topic 1	Topic 2	Topic 3
The Badlands	Bar None	Kowloon Tong Cafe
Trigger	Bullitt	Creations
Toad Hall	The Brick Yard	ABC Bakery Cafe
Lookout	Circa	Stonestown Mall
440 Castro	Monagahan's	100% Sweet Cafe

Table 1: Three of the topics inferred using LDA on user check-ins. The first two topics correspond to the Castro and Marina neighborhoods of San Francisco, respectively, while the third topic primarily consists of Asian-themed Places.

Distance of place to user's previous check-ins We use k-means clustering on the user's previous check-ins. We then use as a feature the distance between the place and the closest cluster center for each user. Figure 5 shows the centers k-means produces for a typical user. Empirically, the centers are often associated with neighborhoods in San Francisco, such as the Mission, the Marina, or SOMA.

Miscellaneous characteristics of places We include additional properies of the place: whether the place has been claimed by a business owner and the number of people who have "liked" the place.

Model

Having defined the features, we now use them in a logistic regression model. We construct a balanced dataset of check-ins and non-check-ins and regress on the training set. Table 2 shows a summary of the parameters of the regression.

A number of the features are significant predictors. Unsuprisingly, the strongest predictor is the number of previous check-ins by the user; many people use the product serially and will check into the same venue repeatedly over time (for example, it is not uncommon to find a user that checks into

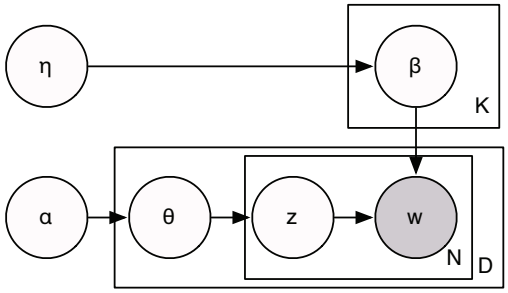


Figure 4: A graphical model depiction of latent Dirichlet allocation. Places are organized into topics β based on their co-occurrence. Users are assigned vectors indicating their participation in each topic θ. Each check-in, w, is associated with a topic for that check-in z.

Figure 5: The result of performing k-means clustering on a typical user's check-ins. Red circles denote the centers of the user's clusters.

his neighborhood bar several times per week). Also significant is the number of check-ins previously made by friends of the user.

Age is also a significant factor which governs overall usage of the product as well as predictions on a per-place basis. Gender, despite the correlations given in the previous section, is not significant after controlling for other factors. The hour of the day of the check-in has a small but significant effect, while the day of week is not predictive.

Some of the topic features computed using LDA also have a significant effect. This means that the clustering is able to find additional predictive signal by correlating across different places.

We then apply our regression model to make predictions on the test set. Figure 6 shows the precision/recall curve on the test set. The model is able to accurately predict check-ins on the test set, achieving approximately 90% precision at 60% recall.

Predicting How People Will Respond

Game-based location services such as Foursquare and Gowalla incentivize users to check in using game mechanics. On these services, users earn badges after accumulating

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-0.4236	0.0878	0.0000
Hour of day	-0.0055	0.0023	0.0161
Gender = male	0.0207	0.0229	0.3675
Age	0.0051	0.0010	0.0000
# of check-ins	-0.0001	0.0000	0.0040
# of check-ins by user	2.2030	0.0890	0.0000
# of check-ins by gender	-0.0000	0.0001	0.5481
# of check-ins by age	-0.0020	0.0003	0.0000
# of check-ins by friends	0.6532	0.0121	0.0000
# of topic 1 check-ins	-0.0101	0.0010	0.0000
# of topic 2 check-ins	-0.0022	0.0021	0.2945
# of topic 3 check-ins	-0.0076	0.0015	0.0000
# of topic 4 check-ins	-0.0087	0.0017	0.0000
# of topic 5 check-ins	-0.0004	0.0015	0.7667
Distance to user cluster	-0.0007	0.0004	0.0587
Distance to user	0.0000	0.0000	0.0000
Day = Monday	-0.0276	0.0476	0.5629
Day = Tuesday	-0.0493	0.0469	0.2936
Day = Wednesday	0.0020	0.0450	0.9649
Day = Thursday	0.0472	0.0436	0.2789
Day = Saturday	-0.0024	0.0374	0.9479
Day = Sunday	-0.0246	0.0401	0.5390
Owned page	-0.0166	0.0073	0.0224
# fans	0.0000	0.0000	0.4160

Table 2: Summary of the logistic regression model to predict check-ins.

a certain number of check-ins or gain special recognition if they have more check-ins at a particular location than all other users.

In contrast, Facebook Places provides no game-based incentives to check in at the time of this writing: there are no badges, no public leaderboards, and no visible acknowledgement of a user being particularly active at any specific place. Other location-based services with specific incentives, such as Facebook Deals, do encourage users to share their locations, but such programs are in their infancy and currently have low participation rates among users and businesses. Instead, participation on Facebook Places is motivated primarily by users' active intention to share their location with friends; their check-in stories appear on their friends' News Feeds and may receive likes and/or comments.

Below we explore how people will respond to users' check-in data via News Feed. Being able to identify factors that increase liking and commenting may yield improvements in how stories are ranked and increasing the probability of feedback and interaction.

In the sequel we build a model to predict whether or not someone will respond to a check-in story in their News Feed. We will use the term *actor* to denote the individual performing the check-in (and thus appearing in others' fields) and *user* to denote the individual seeing, and potentially liking or commenting on the story. We use on distance features for our model, derived from five (sets of) locations associated with the event:

- The usual location of the user (a proprietary determination of where she typically uses Facebook)

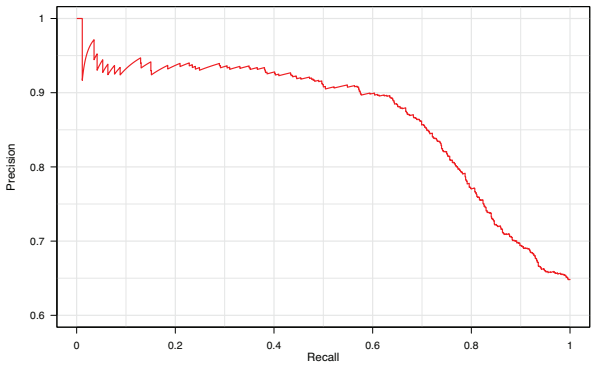


Figure 6: Precision/recall of the logistic model over the test period. Even at 60% recall the model is able to predict check-ins with 90% accuracy.

- The usual location of the actor.
- The location of the check-in.
- The clusters associated with each user (found using k-means and described in the previous section).
- The clusters associated with each actor.

The pairwise distances between these locations yields 10 features for our model. These features are then used in two regression models — one to predict "likes" on the check-in story, and one to predict comments on the check-in story. Two-thirds of the data were randomly split into a training set with the remainder held out as a test set. The result of these regressions are summarized in Table 3.

None of the predictors, except the distance between the user and the actor, were significant for predicting likes. In contrast, many of the features are non-trivially significant for predicting comments. In particular, if the check-in is near a user cluster — that is, if the check-in occurs near the center of an area the user typically frequents — then there is a higher probability of a comment. Similarly, if the actor is near the user, either measured by home location or nearest cluster, then the likelihood of a comment goes up dramatically.

After training these models, we apply them to make predictions on our held-out test data. Because both likes and comments are extremely sparse, this is a particularly difficult prediction problem. The precision/recall curves of these prediction tasks are shown in Figure 7. Commensurate with the previous regression results, comments are better predicted than likes, with the like prediction model barely registering above random guessing.

We can better understand why performance is so poor by plotting how response rates vary with feature values. Figure 8 shows how the feedback rate, i.e. the probability of liking or commenting conditioned on impression, varies as a function of the distance between the two locations given in each caption.

Response rates are highest when the check-in event happens near one of the user's cluster means. At the same time, response rates are strongly peaked when actors are far from

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-3.18	0.08	0.00
(user,actor)	0.05	0.02	0.02
(user,checkin)	0.04	0.04	0.35
(actor,checkin)	0.02	0.03	0.58
(actor,user cluster)	-0.03	0.03	0.43
(user,actor cluster)	0.00	0.04	0.98
(user,user cluster)	-0.00	0.03	0.96
(actor,actor cluster)	-0.00	0.04	0.97
(actor cluster,user cluster)	-0.20	0.04	0.00
(user cluster,checkin)	-0.05	0.03	0.08
(actor cluster,checkin)	0.07	0.04	0.10

(a) likes

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-2.61	0.07	0.00
(user,actor)	-0.07	0.02	0.00
(user,checkin)	0.01	0.04	0.83
(actor,checkin)	0.03	0.03	0.45
(actor,user cluster)	0.10	0.03	0.00
(user,actor cluster)	0.11	0.04	0.01
(user,user cluster)	-0.07	0.03	0.03
(actor,actor cluster)	-0.06	0.04	0.18
(actor cluster,user cluster)	-0.36	0.04	0.00
(user cluster,checkin)	-0.09	0.03	0.00
(actor cluster,checkin)	0.12	0.04	0.00

(b) comments

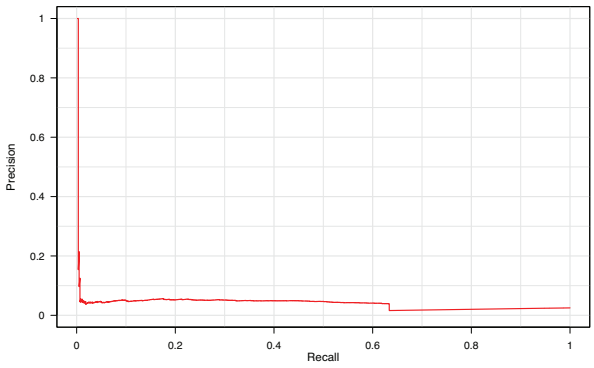
Table 3: Summary of logistic regression models to predict likes and comments on check-in stories in News Feed. Each feature consists of the distance represented by the two locations in parentheses.

their usual location. Note that the peak distance is roughly the width of the United States. Users are clearly responding to check-ins that indicate that the actor is travelling. At the same time, note the curvilinear relationship between actor and user cluster. This plot combines the behavior of the previous two plots. When the check-in is near the user, naturally the actor must also be near the check-in and so the probability of response is high at the lower end of the x-axis. But as we observe in plot (c), when the actor is far from the user (e.g. a nearby friend is travelling), the response rate again increases.

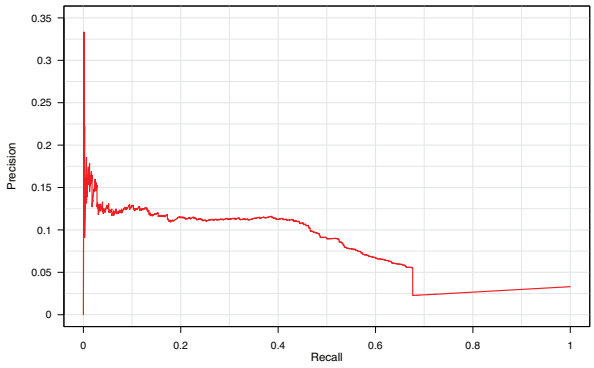
Another dimension which gives insight into this behavior is looking at the response rate for each of the words appearing in the name of a place. This can be thought of as creating a naïve Bayes classifier for likes and comments based on a bag-of-words of the place name. Table 4 shows the words most strongly correlated with likes (left) and comments (right). Responses are liable to be high when the check-in is nearby (such as a hospital or gym) or when the check-in is far (such as at an airport or amusement park).

Predicting Friendship with Check-in Data

Although services like Facebook Places have extremely rich data about its userbase such as demographics and other features used in the previous analyses, one may wonder how useful location-based data is for other location- based appli-



(a) likes



(b) comments

Figure 7: Precision/recall of the logistic model over the test period. Comments are better precited than links; the prediction of likes is barely better than random guessing.

cations that may not have as much data about their users. Perhaps homophily of check-ins is sufficient to derive some useful signal.

To investigate this, we examine the correlation between check-in data and friendship. We take all tuples of (*place*, *actor1*, *actor2*) where both *actor1* and *actor2* have checked into *place* at least once; from this data, we construct a dataset of 794,543 rows with the constraint that half of the actor pairs are friends and half are not friends. We take 2/3 of this data for training data, maintaining the 50% friend pair criterion, and attempt to segregate friend vs. non-friend rows given only the check-in data.

We construct a logistic regression with the following regressors:

- total number of check-ins for the POI

likes	comments
disneyland	hospital
fitness	medical
in-n-out	airport
disney	center

Table 4: Top words predictive of likes and comments.

- number of check-ins by the first actor
- number of check-ins by the second actor

In order to avoid bias, we use only globally-visible places (i.e. places visible only to the place's creator and her friends, such as places like "Home", are not included here). Furthermore, tagged check-ins are not included, so if Alice tags her friend Bob along with the check-in, Bob's check-in is not used in this dataset.

Despite these self-imposed barriers, we still have significant predictive power when predicting friendship, as seen in Table 5.

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	5.520e-01	9.923e-04	0.0000
# total check-ins	-1.444e-05	5.809e-08	0.0000
# actor1 check-ins	2.280e-02	2.864e-04	0.0000
# actor2 check-ins	2.287e-02	2.879e-04	0.0000

Table 5: Logistic regression to predict friendship given only check-in data.

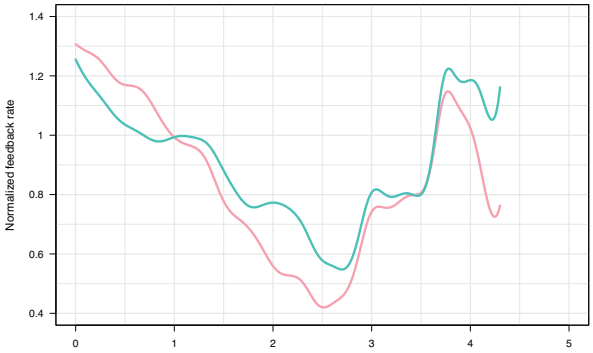
	0	1
0	98084	39337
1	34338	93085

Table 6: Confusion matrix for predicting friendship. 1 denotes a pair of actors that are friends.

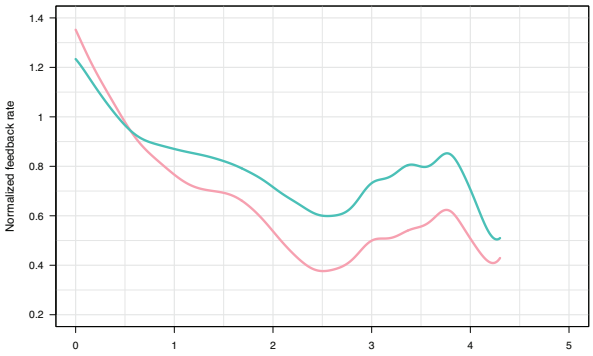
All else equal, each additional check-in by the two actors increases the likelihood that they are friends by approximately $e^{0.0228} - 1 = 2.3\%$. Furthermore, after determining the optimal separation point from the training data, we can predict friendship with 72.18% accuracy on the held-out test set. Table 6 shows the confusion matrix for this prediction. These results suggest that co-check-ins would be useful for suggesting friendship between two Facebook users (for example, in an application like People You May Know) who frequently check into the same places.

Conclusions and Future Work

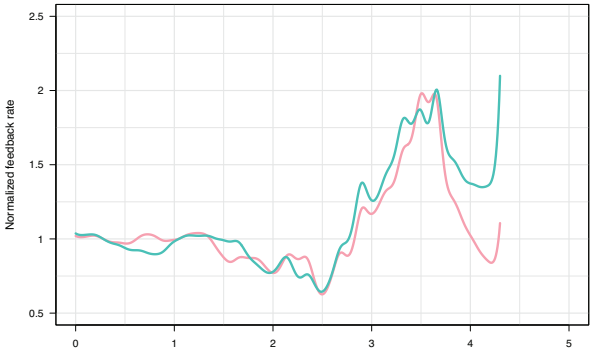
This paper presents several analyses using data collected from Facebook Places. Using a dataset of check-in and POI data from San Francisco, California collected between August 2010 and January 2011, we develop models that predict where users will check in, how their friends will respond, and whether their actions infer friendship. We are able to achieve 90% precision and 60% recall when predicting check-ins on our held-out test set; the most significant predictors are previous check-ins and check-ins by friends. When predicting feedback on check-in stories, we find that the physical distance between the viewer and actor (the one who checks in) is the only predictive feature of likes, but we find several significant predictors of comments on check-in stories. Finally, we find that check-in data shows strong homophily; pairs of users that check into POIs frequently are much more likely to be friends of each other, even after



(a) (actor,user cluster)



(b) (user cluster,checkin)



(c) (actor cluster,actor)

Figure 8: Plots of the response rate (probability of liking or commenting conditioned on impression) as a function of the logarithm of the distance (km). Likes are in blue and comments are in red.

removing tagged checkins and using only globally-visible POIs that are open to all users, not just friends.

Future work will refine and extend the features used in our models. For example, with better category information for POIs (e.g. airports vs. stadiums vs. bars) we can likely improve prediction performance — we have anecdotally seen that users' check-ins are frequently clustered by type. Furthermore, we will broaden our analyses and examine other regions beyond San Francisco; we may compare results of different cities across the United States and across the world to see whether we can replicate these results in regions where technological savviness, privacy attitudes, and density of both users and POIs may be substantially different. Since location-based services such as Facebook Places still have yet to reach maturity, this data will also be useful in analyses beyond the scope of this paper such as analyzing privacy settings in relation to check-ins and predicting the likelihood of users becoming active users of location-based services.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Bonato, B.; Janssen, J.; and Pralat, P. 2010. A geometric model for on-line social networks. In *3rd Workshop on Online Social Networks*.

Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. In *Proceedings of the National Academy of Sciences 107: 22436-22441*.

Cranshaw, J., and Yano, T. 2010. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Lin, J.; Xiang, G.; Hong, J.; and Sadeh, N. 2010. Modeling people's place naming preferences in location sharing. In *12th ACM International Conference on Ubiquitous Computing*.

Scellato, S.; Mascolo, C.; Musolesi, M.; and Latora, V. 2010. Distance matters: Geo-social metrics for online social networks. In *3rd Workshop on Online Social Networks*.

Tang, K. P.; Lin, J.; Hong, J.; Siewiorek, D.; and Sadeh, N. 2010. Rethinking location sharing: Exploring the implications of social-driven vs. purpose-driven location sharing. In *12th ACM International Conference on Ubiquitous Computing*.