A generative model of urban activities: simulating a population

Rory McGrath
University Of California, Berkeley
rorymcgrath@berkeley.edu

Alexey Pozdnukhov
University Of California, Berkeley
alexeip@berkeley.edu

ABSTRACT

This work presents a data-driven generative model of a social network in a city. It takes both mobility and social context into account by building a generative process of establishing social connections that is governed by latent profiles of city dwellers. These profiles influence both the community structure of the network as well as preferences in destination choices. Therefore, the model captures an interplay between mobility and the geography of the social network, and can be used to infer social connections from the revealed mobility pattern and vice versa. The model is capable of inferring population profiles and community structure of urban areas from partly observed social network and mobility traces, and could then be applied to simulate a representative sample of the entire population. The results are verified in terms of descriptive statistics of the network structure as well as on the out-of-sample subset of real data collected via Twitter API in San Francisco, CA.

General Terms

Algorithms, Generative modeling, Networks

Keywords

Generative Model, Social Network, Prediction, Machine Learning

1. INTRODUCTION

Currently data availably from check-in services only tells half a story. This is not a fault of the service but, besides behavioral and demographic bias, arises from users being influence by people not observed by the system. These latent forces can cause observed users to appear unpredictable or act in peculiar ways making reliable prediction and community detection somewhat challenging. If, however, these latent individuals were observed the patterns and motivation behind all users actions would become more apparent making it easier to both predict and apply semantic meaning to events. For this work we look at generating this complete dataset for all inhabitants in a city.

From [3] we know that a city is made up of various communities of people. In order to generate realistic tweets we first need to simulate this social network. Extending the work from [1] we expand an existing twitter social network taking into account both preferential attachment and spatial preferences.

Check-ins are created for the individuals using a modified Link-PLSA-LDA algorithm. Probabilities for individuals visiting a location are calculated by considering the radiation model, an individuals personal preferences and their friends preferences.

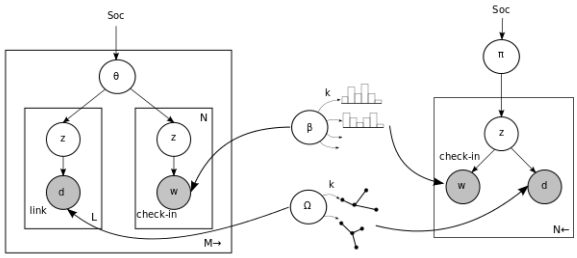

Figure 1: Generative model for check-ins.

Tuning the parameters in our algorithm using real world data our model generates realistic check-ins for all users in a city. These check-ins are created each time a user visits a location or venue with the users probability of visiting a location being determined by a weighted combination of the communities they are members of, their standing in each community, their loyalty to locations and the travel distance required to reach the venue.

2. ALGORITHM OVERVIEW

In order to generate check-ins the algorithm follows a set of elegant and logical rules. Using these rules and training on real-world data a set of realistic check-ins are generated for all individuals, both observed and latent, using a variate of the Link-PLSA-LDA model. Using the terminology of PLSA our individuals are 'documents' with the location of their check-ins being their 'words'. People connect to or 'cite' each other through social links/friendships. Our model is presented in figure 1 with the following parameters being used:

UrbCompâĂŹ14, August 24, 2014, New York, NY, USA location

$\Omega_{d'}$ the probability of a friendship to person d'.
β the probability of a location with respect to a topic.
π the probability of topic for the set of people.
θ the topic distribution of a person.

The code used in the generation process is describe in pseudo code (algorithm 1). Here check-ins are generated for each individual based on their topic distribution and friend connections.

Before creating a check-in a current representative sample of the population is obtained. This populated is created from the set of current active twitter users in San Francisco. For each person their home and work locations are inferred from geo-referenced tweets. A distribution over topics θ_p is also assigned to each individual. This θ_p, generated using census data, shows an individuals likelihood to participate in each topic thus representing their interest levels. These topics can encompass activities such as socializing, leisure, shopping. New individuals are then generated with home locations sampled from a distribution over residential areas. The initial twitter social network is expanded to include these individuals using a combination of both preferential attachment and spatial convenience.

Initial probabilities of an individual visiting each location for each topic are calculated using a radiation model trained on venues in San Francisco. The population of San Francisco and the number of venues is immense, to account for this we created a radiation model which shows an individuals probability of visiting an area at the granularity of the tract level as defined by the census.

In order to generate a check-in for a person we first find their potential friends. The probability of a friendship p_{friend} is calculated using product of three components:

$$p_{friend} = s_{pf} \times d_{pf} \times deg_{pf} \quad (1)$$

The first component, s_{pf}, deals with the similarity between person p's topics of interests θ_p and person f's topics θ_f, calculated using the cosine similarity. The second component, d_{pf}, is the distance decay between person p's home and person f's home. Intuitively here we see that the closer people live together the more likely a friendship. The final component, deg_{pf}, is the friends f degree in the social network, the more popular the individual the more likely a friendship. Using these values and normalizing, the probability distribution of all potential friends F is created for a person p.

Once a friend is found the topic of the check-in is determined. Two types of check-ins are possible, social and non-social. If a check-in is non-social the topic is found by sampling from θ_f. If the check-in is social the topic is found by sampling from θ_f and θ_p, reflecting a topic that is of interest to both parties.

With the friend and topic determined the final part of the algorithm deals with finding the location of the check-in. The probability of each venue being chosen is calculated using the equation:

$$p_{checkin} = p(v|s=0) \times p(s=0) + p(v|s=1) \times p(s=1) \quad (2)$$

Here $p(s)$ represents the probability of a social check-in. For non-social check-ins the probability of location relies solely on the individual. This probability $p(v|s=0)$ is calculated using the radiation model returning the probability of each venue v for the given topic t.

For social check-ins the choice of location is influenced by both parties using a mixture of the radiation models for person p and f and the probability of a friendship (friendship strength) between p and f. The results are normalized accordingly.

$$p(v|s=1) = rad_{pt} + p_{friend} \times rad_{ft} \quad (3)$$

Check-in events are then assigned to the friend. By applying the check-in to the friend the power law distribution of check-ins that is observed in real world data is maintained. It can be seen that the higher the chances of p and f being friends the greater the influence p has on f. This process is then repeated.

3. FRAMEWORK

The algorithm proposed in this work consists of two core components, the social network of the users and the spatial network of the venues. For this work it was decided to use the Neo4j database to handle the social network and MongoDB to manage the spatial network.

Neo4j is a scalable robust native graph database. Our users are represented as nodes in a graph with the edges representing friendships. This highly scalable database allows us to work with populations of thousands and scales to include populations of millions with optimized queries to calculation an individuals degree, friends and friends of friends etc.

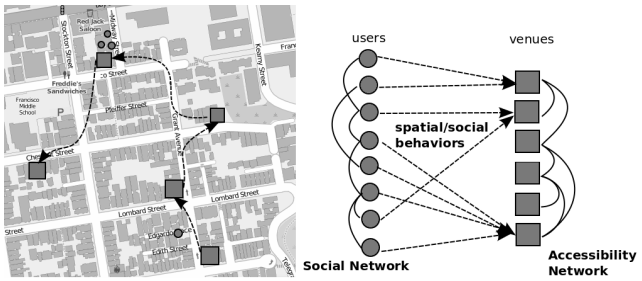

Figure 2: The framework of algorithm. Here individuals are connected via a social network and locations are connected via accessibility. People frequent locations based on spatial proximity or social recommendations

Venues were saved as documents in MongoDB which included a spatial component and other meta data. MongoDB is a scalable document-orientated database and similar to Neo4j allows us to work with thousands of venues but easily scales to accommodate millions offering various index and query mechanisms to handle geospatial data. Venues are connected by an accessibility network capturing the most probable next locations given the current venue.

Algorithm 1 pseudo code

θ_p = distribution over topics for person p.
rad_{pt} = radiation model for person p and topic t
P = generate all people in population each with θ_p
G = initial social network
V_t = all venues for topic t
for each person $p \in P$ **do**
  for each friend $f \in P$ and $f \neq p$ **do**
    $p_{friend} \approx s_{pf} \times d_{pf} \times deg_{pf}$
    $\rightarrow s_{pf}$ is similarity of θ_p and θ_f
    $\rightarrow d_{pf}$ is the distance decay from p_{home} and f_{home}
    $\rightarrow deg_f$ is the friends f degree in G
  end for
  create friend distribution F
  sample from F to find friend f
  sample from θ_f to find topic t
  for each venue $v \in V_t$ **do**
    $p_{checkin}$calculate probability of check-in
    $p_{checkin} = p(v|s=0) \times p(s=0) + p(v|s=1) \times p(s=1)$
    $\rightarrow p(s)$ = probability of social check-in
    $\rightarrow p(v|s=0) = rad_{ft}$
    $\rightarrow p(v|s=1) = rad_{pt} + p_{freind} \times rad_{ft}$
  end for
  create check-in distribution C
  sample from C to get check-in
  create link $\propto p_{friend}$
  assign check-in to f
end for
repeat until some metric for convergence

Users and venues are connected via spatial and social behaviors. Spatial behaviors affecting a persons choice of venue include their willingness to travel, their sensitivity to travel time and the accessibility of the venue. The social behaviors affecting the choice of venue include the persons willingness to pay, or value for money of the location, the rating of the location and the desire of friends to visit this location. Venues are connected through an accessibility or spatial network showing what locations are reasonable accessible given the current location of the individual. This accessibility is given by spatial proximity and travel time.

4. DATA

For this work we focused on San Francisco and the Bay area. Twitter data was collected for a period of one month resulting in 1.3 million tweets and 12,000 unique users. We fixed the venues of interest to be restaurants. Utilizing the Google location api restaurants in San Francisco and the surrounding Bay Area were located.

4.1 Social network generation

Each unique Twitter user with sufficient tweets (more than ten on weekdays) was identified and added as a node in the Neo4j database.

Utilizing the work from [2] and [5]each users home and work locations were inferred. [2] built on the observation that people show strong periodic behavior throughout certain periods of the day. This behavior alternates between primary ,"home", and secondary, "work", locations on weekdays. This distribution was modeled with a truncated Gaussian distribution parameterized by the time of the day:

$$N_H(t) = \frac{P_{c_H}}{\sqrt{2\pi\sigma_H^2}} exp[-(\frac{\pi}{12})^2 \frac{(t-\tau_H)^2}{2\sigma_H^2}]$$
$$N_W(t) = \frac{P_{c_W}}{\sqrt{2\pi\sigma_W^2}} exp[-(\frac{\pi}{12})^2 \frac{(t-\tau_W)^2}{2\sigma_W^2}]$$

and

$$P[c_u(t)=H] = \frac{N_H(t)}{N_H(t)+N_W(t)}$$
$$P[c_u(t)=W] = \frac{N_W(t)}{N_H(t)+N_W(t)}$$

Here τ_H is the average time of the day when a user is at home and σ_H is the variance in time of day. P_{c_H} is the time-independent probability that any given check-in was generated by the "home" state. The same holds true for the work state. Of the 12,000 unique users found 1,000 had sufficient data to generate reasonable parameter estimates for their home and work locations.

Using these individuals an initial social network for the city was created. Friend connects were determined by looking at all reciprocal followers on twitter. Here a friendship was defined by a mutual following on twitter. These friends were used to populate the social network in Neo4j and were assigned an appropriate θ vector. Taking this as the initial social network, the network was then expanded.

New individuals were generated with home locations sampled from the distribution of residential areas obtained from census data. These were added to the network using a combination of preferential attachment and spatial proximity.

4.2 Spatial Choice model

The venues of interest for this work were fixed to restaurants in San Francisco and were obtained using the Google Locations api. In order to find the probability of an individual checking into a venue the radiation model was used. The radiation model is defined by the equation:

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \quad (4)$$

Where T_{ij} represents the commuting fluxes, m_i represents the population at the source i, n_j represents the population at destination j, s_{ij} represents the total population in a circle of radius r_{ij} centered at i, and T_i is the total number of commuters who start their journey at i.[4]

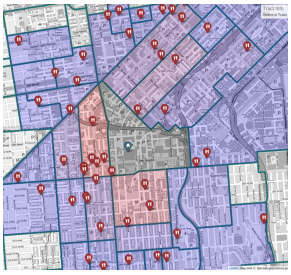

Figure 3: Radiation model for a user based on their home location

5. RESULTS

In this section we combine both the social and spatial components of our framework and generate check-ins as defined by the algorithm above. The results presented below include the distribution of check-ins and the distribution of the degree in the generated social network. Using our generative process check-ins were created for the inhabitants of San Francisco. Link-PLSA-LDA was preformed on the generated check-ins to recover any latent topics that appeared in our generated data.

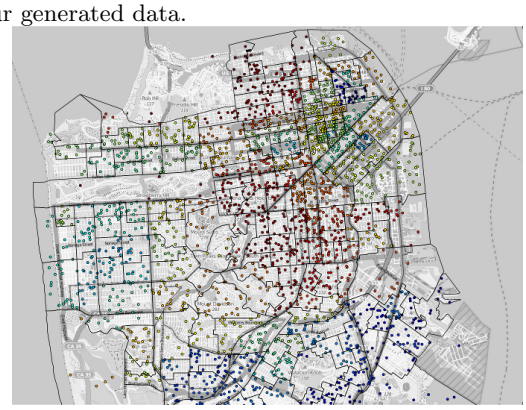

Figure 4: Distribution of topic 1

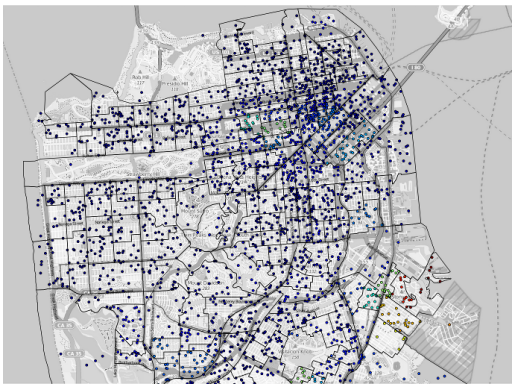

Figure 5: Distribution of topic 2

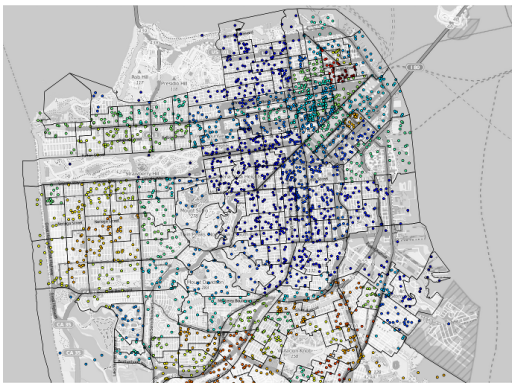

Figure 6: Distribution of topic 3

The distribution of actual topics are given above. In order to recover any communities Link-PLSA-LDA we preformed for 3 topics and results are given below. Here we found the topics returned adhered to the topic distributions shown above.

Figure 7: Estimated distribution of topic 1

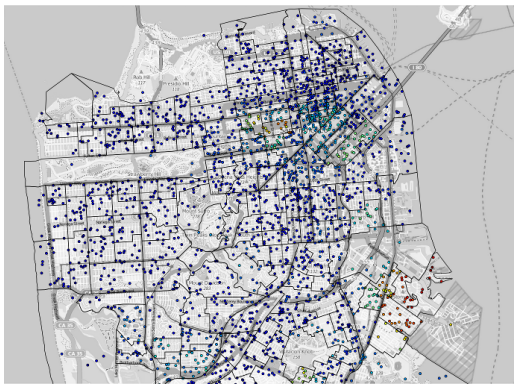

Figure 8: Estimated distribution of topic 2

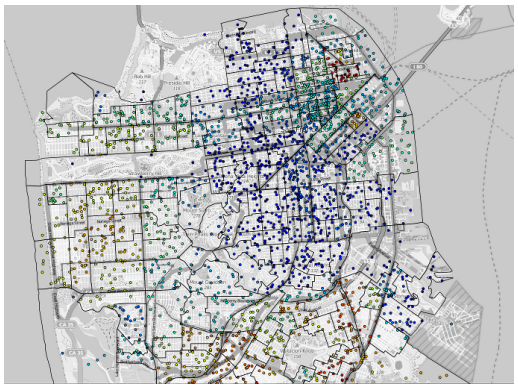

Figure 9: Estimated distribution of topic 3

Looking at the correlation plots between estimated and real topics we can see that out proposed algorithm does a good job at simulating users check-in interests.

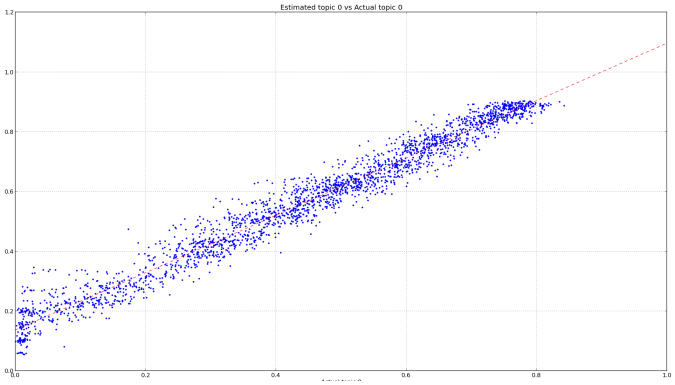

Figure 10: Correlation of topic 1

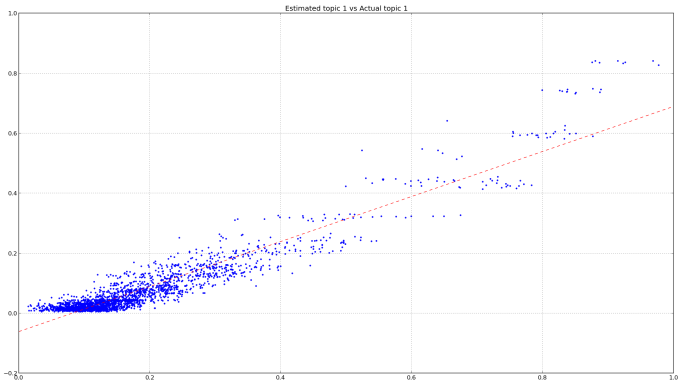

Figure 11: Correlation of topic 2

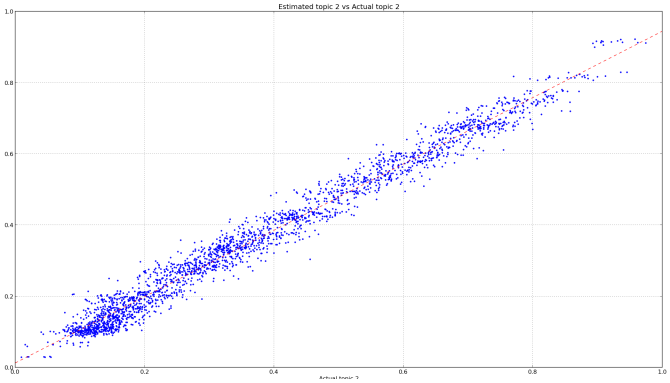

Figure 12: Correlation of topic 3

Next we look at the distribution of the tweets generated per individual and the distribution of the degree in the social network. Again we see that the values obtained are consistent with prior expectations of social networks and observed behaviors of twitter users. Here the generative process captures the property that some users are extremely active when using the service while others take a more passive role.

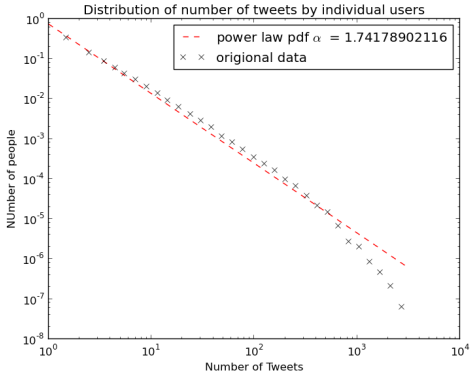

Figure 13: Distribution of generated tweets for each individual

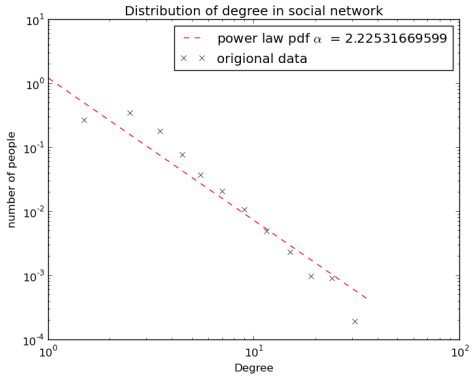

Figure 14: Distribution of degree in the generated social network

6. SUMMARY & CONCLUSION

The work here presents a generative model and robust framework for the creation of realistic check-ins for a cities inhabitants which arise from friend interactions and an individuals preferences of locations.

Utilizing real-world data from San Francisco, check-ins and a social network were created for the cities inhabitants which looked to account for the latent forces that cause observed users to appear unpredictable or act in peculiar ways.

Currently the parameters of this generative process are obtain from census data or through experimentation. Future work includes incorporating a discriminative model to estimate new parameters based on real world data and iteratively converge on the true value of these estimated parameters. This would allow form a more accurate representation of social networks and check-ins for cities.

7. REFERENCES

[1] Barabasi, A. L., and Albert, r. Emergence of scaling in random networks. *Science 286*, 5439 (1999), 509–512.

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

[3] Lawlor, A., Coffey, C., McGrath, R., and Podzdnoukhov, A. Stratification structure of urban

habitats, PURBA - 2012.

[4] Simini, F., Gonzalez, C. M., Maritan, A., and Barabasi, A. L. A universal model for mobility and migration patterns. *Nature 484*, 10856 (2012), 96–100.

[5] Tarasov, A., Kling, A., and Pozdnoukhov, A. Prediction of user location using the radiation model and social check-ins, UrbComp - 2013.