Final Report
Evaluating Social Networks as a Medium of Propagation for Real-Time/Location-Based News

Mehmet Ozan Kabak, Group Number: 45

December 10, 2012

Abstract

This work is concerned with modeling the propagation of real-time/location-based news over social networks. We start out by presenting a brief introduction to the problem of news propagation in the real-time/location-based setting. Then, we give a short literature survey and also briefly talk about the existing work in modeling the mentioned problem. In the third section, we propose a new model to evaluate social networks as a medium of propagation for real-time/location-based news. The model is also partially validated using tweet data about a real-world news story. We then explain our simulation framework and our data set as well as the required preprocessing for the data. Finally, we present some results and conclusions.

1 Introduction: Propagation of real-time/location-based news

Social networks have been under the spotlight in the recent years for their potential (and current) use in news propagation. Empirical observations regarding the role of prevailing social networks like Facebook and Twitter in the propagation of big-impact news (e.g. news about the "Occupy Wall Street" demonstrations, developments regarding the Arab Spring) have sparked a lot of interest about the topic. Although many studies have been conducted regarding the cascading behavior in networks, few (if any) models (and results) are available about the general dynamics of news propagation that captures the temporal and spatial aspects of the problem.

Even though many qualitative arguments are commonly used to argue the efficacy of social networks as a medium of real-time/location-based news propagation, a well-established model that is able to produce quantitative results is still lacking.

2 Short Survey of Existing Work

As mentioned in the previous section, we were not able to find a work that appropriately models both the temporal aspects and the spatial aspects of the process of news propagation

over a network. In the "Reaction Paper/Project Proposal" report, we discussed two papers ([1], [2]) that focused on accurately modeling the temporal aspects of information propagation in a social network. In [1], a variation of the Independent Cascade Model is used to analyze "long-running-type" and "spike-type" topic activity in the blogosphere. In [2], a novel model called "Linear Influence Model" is presented. This model tries to estimate the global diffusion rate by analyzing the cascade at the individual node level. Using this model, the authors were able to explain many aspects of the (temporal) dynamics of information diffusion.

Contrary to the previous two papers, the authors examine a location-based social network in [3]. This paper develops a model of node mobility by analyzing the check-in data of a real life location-based social network. Even though the paper is not related to information propagation, it is still somewhat relevant since it constructs a movement model that is both time and location dependent.

3 Proposed Model

In this section we elaborate on our model that was first proposed in the "Reaction Paper/Project Proposal" report. There are two main entities in our model: news stories and users. A complete description of the representations of these entities and their interaction model is presented below in greater detail.

3.1 Modeling the news stories

A news story n_i is represented by a quintuple with the following five fields:

- Geographical coordinates of the place the news story originates from; denoted by $d(n_i)$.
- Origination time of the news story; denoted by $t(n_i)$.
- A coefficient modeling the geographical locality of the news story; denoted by $\alpha(n_i)$.
- A coefficient modeling the temporal locality of the news story; denoted by $\beta(n_i)$.
- A coefficient modeling intrinsic "attractiveness" of the news story; denoted by $p(n_i)$.

3.2 Modeling the users

A user u_i is represented by a tuple with the following fields:

- Location of the user; denoted by $d(u_i)$.
- The rate at which the user becomes active; denoted by $\lambda(u_i)$.

In the proposed model we assume that users become active (i.e. may take part in the information diffusion process) only at discrete times. The times when a particular user becomes active is modeled by a Poisson process, whose rate is given by $\lambda(u_i)$.

3.3 User-news interaction model

We assume that there are two ways a user u_i can interact with a news story n_i that reaches him/her at time t. The first interaction is an "observation": The user either finds the news story interesting or not. The probability of the former outcome is modeled by:

$$\Pr\{u_i \text{ finds } n_i \text{ interesting}\} = f\left(p(n_i), \alpha(n_i)\left|d(u_i) - d(n_i)\right|, \beta(n_i)(t - t(n_i))\right) \tag{1}$$

where f is a function that will be discussed in greater detail in subsequent sections.

The second possible interaction is the "transmission" (e.g. resharing, retweeting) of the news story by the user. When a user transmits a news story, the story reaches the neighbors of the user once they become active. Note that this interaction is conditional on the first interaction; i.e. the user has to find the news story interesting for a possible "transmission" to occur. The probability of this event is modeled by:

$$\Pr\{u_i \text{ transmits } n_i | u_i \text{ finds } n_i \text{ interesting}\} = g\left(\Pr\{u_i \text{ finds } n_i \text{ interesting}\}\right) \tag{2}$$

where g is a function that will be discussed in greater detail in subsequent sections.

3.4 Interpreting the model

Assume that f is a function that is monotonically increasing in its first argument and monotonically decreasing in its second and third arguments. With this choice, we see that the probability of a user finding a news story interesting decreases as the origin of the news story and the location of the user get farther apart. Similarly, the probability is also inversely proportional to the temporal distance between the time the news story reaches the user and the time it originated. Also, we see that by varying the coefficients α and β, we can model locally or globally relevant news stories as well as long-lasting or short-lived news stories. Now consider the following particular choice:

$$f(p, d, t) = p \times \exp(-d) \times \exp(-t). \tag{3}$$

In this scenario, the parameter β has a very intuitive meaning: It is nothing but the reciprocal of the half-life of the news story (within a multiplicative factor of $\ln 2$). Similarly, the parameter α becomes the half-distance of the news story (again within a multiplicative factor of $\ln 2$). Note that half-distance is simply the spatial analog of half-life. Interestingly, the parameter p also takes on a very intuitive meaning in this scenario; i.e. it becomes the probability of a user finding the news story interesting if he/she were "witnessing" the story right at its origination location/time. We will use this particular form (i.e. equation (3)) in the context of this project as it results in easily interpretable model parameters, .

Now consider the function g: The specifics of the "transmission" process is modeled by this function. For example, $g(x) = 1$ would mean that users would be "transmitting" everything they liked. As another example, consider $g(x) = x$. This would imply that a

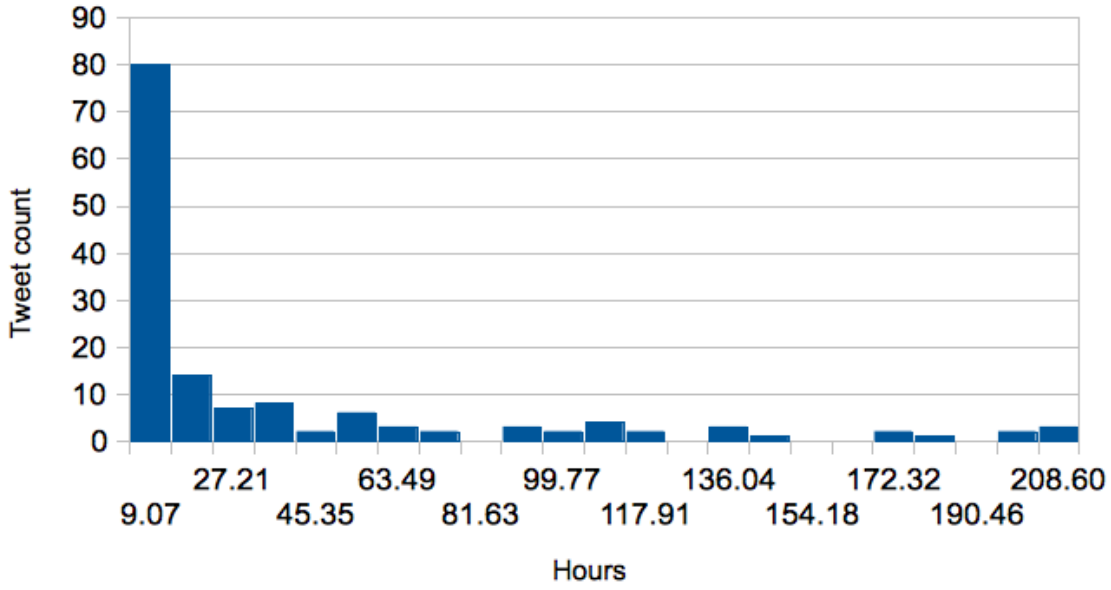

Figure 1: Histogram of the number of tweets about the Los Angeles Lakers - Orlando Magic NBA Final Game on 06/14/2009 at 8pm.

news story would need at least 0.7 probability of being found interesting for achieving a "transmission" probability of 0.5.

Modeling the "transmission" in a two-step manner gives us two explicit knobs that enables us to reason about the cascading behavior under different assumptions (e.g. about virality) with a simple change to one of the two knobs. Over the course of this project, we have experimented with a few different vitality models. However, for the results presented in this report, we fixed $g(x) = x$. The main motivation behind this choice is to reflect the idea that users are more likely to "transmit" things they like more.

4 Model Validation

There are unfortunately no empirical data about the propagation of news stories in a location-based context. However, there exists such data that have embedded temporal information. Hence, temporal aspects of the presented model can be validated.

For this work, we chose to utilize tweet data for the purposes of model validation. Results for a particular news story is presented here. The news item of choice is the last game of NBA season 2009, which took place between Los Angeles Lakers and Orlango Magic, on 06/14/2009 at 8pm. Figure 1 presents the histogram of the tweet count concerning this event where the bins correspond to 9 hour wide time intervals. Note that we indeed observe a strong dependence on time, agreeing nicely with the proposed model which hypothesizes a decay-type relationship.

5 Simulation Framework and Dataset

5.1 Simulation Framework

Since our model is a probabilistic one, we used a Monte-Carlo type simulation framework. Prior to the simulations, the user which "creates" the news story was determined via random sampling. During each Monte-Carlo trial, a priority queue of "active" users are maintained. Users were "prioritized" with respect to their wake-up times. Whenever a user wakes up, he probabilistically transmits the news story to his neighbors (e.g. friends, followers). Among these neighbors, the ones who receive the news get added to the queue. Once there are no more users left in the queue, the Monte-Carlo trial terminates.

5.2 Dataset and Required Preprocessing

The biggest dataset with embedded location information in the Stanford Large Network Dataset Collection is the Gowalla dataset. Gowalla was a location-based social networking website where users shared their locations by checking at venues. This dataset was one of the datasets used in [3] to develop node mobility models.

Since location information is only available in the form check-ins, we preprocessed the dataset to infer sensible user locations from the given information. In particular, we used the following method:

- First, we removed all nodes without any check-ins from the network.
- For each of the remaining nodes, we sorted check-ins of each node w.r.t latitudes and longitudes. The median latitude and the median longitude were saved as the location of the node.

The raw dataset consisted of 196591 nodes and 950327 edges with 6442890 check-ins. After preprocessing, we obtained a network of 107092 nodes and 456830 edges where each node was equipped with a latitude/longitude pair.

6 Results

We first present results from two particular simulation runs. In the first simulation, we investigate the propagation of a "small-impact" local news story. To model such a story, we chose an intrinsic attractiveness of 0.75, a half-life of 4 hours and a half-distance of 20 km. The story was created by the user 152334 at the location $(29.7426, -95.4022)$ at time 0. In this simulation, the users of the network were assumed to become active at a rate of 1/(2 hours).

Table 1 presents the results of one particular realization. The following comments are in order:

Node Id	Latitude	Longitude	Probability
152334	29.7426	-95.4022	0.75
10729	29.7426	-95.4022	0.75
67087	29.7426	-95.4017	0.74886
11023	29.7428	-95.4037	0.746161
11166	29.741	-95.4039	0.743644
60185	29.7428	-95.4049	0.743199
63874	29.7426	-95.3987	0.741338
37050	29.7446	-95.4049	0.741231
32681	29.744	-95.4063	0.738996
25125	29.7427	-95.3968	0.736615
152889	29.743	-95.3967	0.736304
90868	29.7479	-95.4016	0.734887
58527	29.743	-95.4091	0.732918
33767	29.7442	-95.4093	0.731925
34274	29.7368	-95.3993	0.731892
55351	29.743	-95.3948	0.731685
3828	29.7485	-95.406	0.730889
62950	29.741	-95.4103	0.729363
125358	29.743	-95.4106	0.729174
121668	29.7468	-95.4096	0.728348

(a) Users with 20 highest interest probabilities

Node Id	Latitude	Longitude	Probability
10729	29.7426	-95.4022	0.75
41585	29.7318	-95.3944	0.713834
68756	29.7344	-95.4176	0.705934
10727	29.7618	-95.4158	0.68755
90631	29.6819	-95.3051	0.502509
10736	29.7837	-95.5455	0.452578
134835	29.5556	-95.3822	0.363504
90634	29.7977	-95.6207	0.35026
131379	29.5745	-95.1476	0.256939
10557	30.2699	-97.7494	0.000228922
2190	32.9393	-96.845	1.3796e-06
854	32.9717	-97.0453	9.57722e-07
307	28.3565	-81.5793	3.34835e-21
4271	44.9816	-93.44	1.7064e-26
11241	37.5466	-77.643	9.00205e-29
880	59.6765	17.526	1.51451e-126
4956	59.3489	17.9159	3.95784e-127
9735	47.05	7.9087	1.44927e-129

(b) Users the news has reached

Table 1: Results from a sample realization for the "small-impact" local news story

- Even though the story propagated to a few users with reasonable interest probabilities, it did not propagate to the "best audience": i.e. the users that are most likely to find the story interesting. This is most probably due to the following reason: If people in the "best audience" are not close (in the graph sense) to the news originator, the "small-impact", local story will probably not survive the decentralized news propagation process until it reaches the destination.
- The story was also transmitted to some users which have virtually zero chance of finding the story interesting. This is an interesting result that matches the "junk" that we observe in our feeds in the prevailing location-based social networking platforms.

In the second simulation, we investigate the propagation of a "big-impact" global news story. For this purpose, we chose an intrinsic attractiveness of 0.95, a half-life of 24 hours and half-distance of 1000 km. The story was created by the user 105860 at the location $(59.9395, 10.7645)$ at time 0. Similar to the first case, the users of the network were assumed to become active at a rate of 1/(2 hours).

Node Id	Latitude	Longitude	Probability
105860	59.9395	10.7645	0.95
189336	59.9403	10.7646	0.949941
42822	59.9382	10.7643	0.949911
98885	59.9385	10.7665	0.9499
174615	59.9375	10.7651	0.949854
148102	59.9392	10.7591	0.9498
173139	59.9422	10.7655	0.949797
123532	59.9365	10.7642	0.949786
56881	59.937	10.7679	0.949781
46461	59.9365	10.766	0.949775
118292	59.9423	10.7685	0.949745
106023	59.9364	10.7679	0.949743
103913	59.9413	10.7705	0.949742
46777	59.9402	10.7718	0.949728
98472	59.9356	10.7654	0.949715
80929	59.9418	10.7713	0.949701
111363	59.9435	10.7631	0.9497
155384	59.9351	10.7655	0.949679
121826	59.9438	10.7619	0.949671
21823	59.935	10.7644	0.94967

(a) Users with 20 highest interest probabilities

Node Id	Latitude	Longitude	Probability
105860	59.9395	10.7645	0.95
105853	59.9348	10.7597	0.949614
97694	59.9341	10.7602	0.949576
105851	59.9332	10.7591	0.949503
17901	59.9324	10.7591	0.949442
62594	59.9303	10.7465	0.949058
146666	59.9274	10.7524	0.949008
111078	59.9234	10.7666	0.948822
17115	59.9235	10.7572	0.948804
17171	59.9255	10.7403	0.948648
100897	59.9245	10.7399	0.948578
97064	59.92	10.7537	0.948525
105858	59.9312	10.7275	0.948513
101372	59.9183	10.7507	0.948371
98886	59.9264	10.7285	0.94837
101371	59.9195	10.74	0.948287
105850	59.9401	10.7169	0.948255
100901	59.9151	10.7698	0.948209
95707	59.918	10.7409	0.948209
67992	59.9165	10.745	0.948177

(b) Same as (a); among the users the news reached

Table 2: Results from a sample realization for the "big-impact" global news story

Table 2 presents the results of one particular realization. We see that the network does a much better job at delivering the news to the right audience. This result is also in good agreement with our empirical observations: Social media is an efficient channel for distributing "big-impact" news.

We also present some statistical results obtained through Monte-Carlo trials. To evaluate the effectiveness of social networks in propagating local-based news, we swept the half-distance of a "hot" news story (i.e. has intrinsic attractiveness 0.95) from 5 km to 2000 km. For each configuration, we measured the sample average of the probability that a "transmission" received by a user is actually found "interesting". As we see from table 1, the absolute value of this number is quite low due to junk "transmissions". However, the relative change in this number gives important information about the change in the effectiveness of the network structure as geographical locality of the propagated news changes. The curve is given in figure 2. We see that the quality of the delivered content is maximum for "global" news whereas the average probability drops almost by an order of magnitude for "local" news. This result is in good agreement with the individual simulation runs pre-

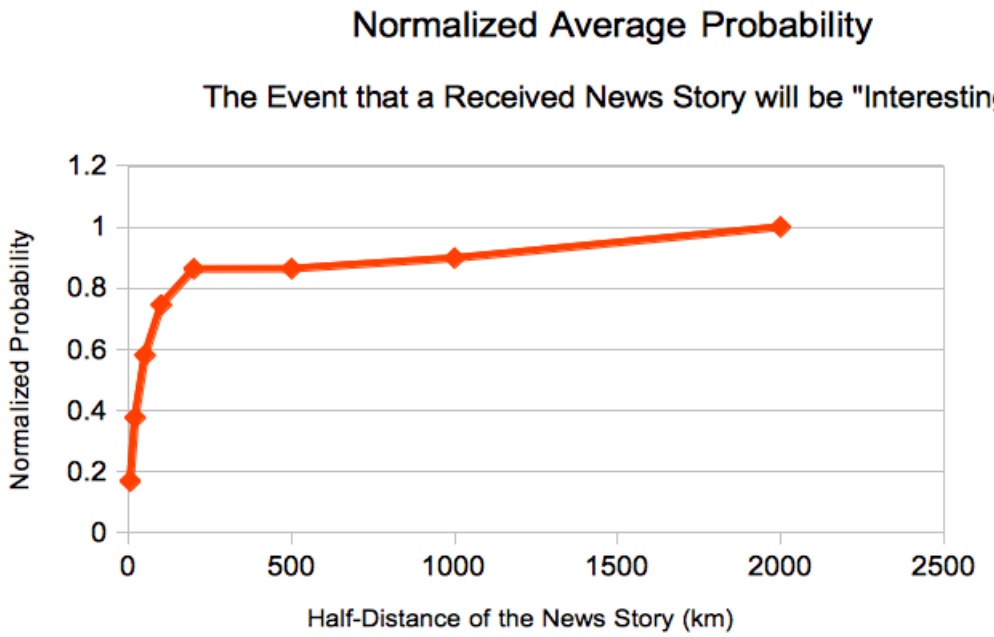

Figure 2: Evaluating effectiveness of a social network for medium of propagation of geographically local news.

sented previously. Finally, note that the curve straightens after a half-distance of 200 km. This is an interesting phenomenon which says that news stories require a certain "globality" (i.e. "geographical independence") in order to effectively propagate in a social network.

7 Conclusions

In this work, we presented a simple and intuitive model for news propagation in a real-time/location-based context. The model was partially verified using real-world data. We then performed simulations to assess the effectiveness of social networks in propagation real-time/location based news. We saw that social networks do a very good job at propagating "big-impact" news but not so much for "small-impact" local news. This is in good agreement with our every-day observations: Global-scope news propagate very fast on social networks whereas it is rare to see local news from your town.

References

[1] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, "Information Diffusion through Blogspace", Proc. International WWW Conference, 2004

[2] J. Yang, J. Leskovec, "Modeling Information Diffusion in Implicit Networks", IEEE International Conference On Data Mining (ICDM), 2010

[3] E. Cho, S. A. Myers, J. Leskovec. "Friendship and Mobility: Friendship and Mobility: User Movement in Location-Based Social Networks" ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2011.