Investigating Temporal Variations in the Twitter Hashtag Graph

Pararth Shah Shantanu Joshi
Stanford University
{pararth, joshi4}@cs.stanford.edu
CS224W: Group 2

December 11, 2013

Abstract

The increasing amount of data shared everyday on social networking sites is a rich source of information about the online as well as the offline world. "Hashtags"[1] are a semi-structured data format which are easy to track, index and aggregate; yet they cover a very diverse range of use since anyone can create and share them. The co-occurrence of hashtags can signify the strength of association between two entities. In this paper, we present a model for computing a similarity score of two hashtags which takes into account the co-occurrence frequency as well as the structure of the hashtag graph. Analyzing the variation of this score over different time intervals gives us a method of identifying the quality of the association between the two hashtags.

1 Introduction

1.1 Motivation

Hashtags are an essential part of the ethos of social networks like Twitter, Instagram and more recently Facebook. They are very similar to free-form text, since anybody can create a hashtag and there is no restriction on its contents. However, hashtags provide context to a piece of online content, and they can be indexed and searched like more structured data formats. Consequently, hashtags reside in a sweet spot between unstructured text and structured data. They can be used to signify products (#nike), events (#worldcup2010), people (#bolt), as well as abstract entities like emotions (#sofunny, #iquit), ideas (#beproud) or movements (#occupywallstreet). The set of hashtags that co-occur consistently makes for an interesting data source on the associations people make between ideas, products, places and even other people.

The rising popularity of knowledge-sharing communities such as Wikipedia and progress in extracting information from text sources on the Web have enabled the automatic construction of very large knowledge bases [1]. DBpedia, Freebase, KnowItAll, ReadTheWeb, and YAGO are some examples of such efforts. These projects provide facts about named entities and relationships between them. They contain millions of entities and relationships, constructed automatically from encyclopedia articles, news content and other Web pages. These structured knowledge bases enable knowledge-centric services like text disambiguation, semantic search, named entity recognition, etc. Prominent examples of how knowledge bases can be harnessed include the Google Knowledge Graph and the IBM Watson question answering system.

We hypothesize that the co-occurrence of hashtags is a significant source of information on implicit relationships between popular topics. Harnessing this largely unexplored information source can provide valuable insights directly applicable to knowledge bases and semantically aware services. However, the very nature of hashtags presents a unique set of advantages and challenges different from that of traditional knowledge sources like news and encyclopedia articles. Due to their crowd-sourced nature, popular hashtags depict what the masses of online users are talking about at a given time. This gives access to the general sentiment on popular issues or recent events, in contrast to news articles which reflect the opinions of the authors, and encyclopedia content which cannot keep up with the latest issues in real-time. On the other hand, this crowd-sourced nature also implies that hashtag data

[1] Words or phrases preceded by a hash or pound sign (#) and used to identify messages on a specific topic.

Total no. of tweets	Approx. 3 billion
Time span	Jan 2011
Dataset size (compressed)	855 GB
No. of hashtag mentions	465,832,570
No. of unique hashtags	24,792,561
No. of unique hashtag pairs occurring in at least 1 tweet	35,179,966

Table 1: Dataset statistics.

is prone to very high levels of noise due to large numbers of personal, ambiguous or misspelled hashtags that do not add any meaningful information. A model of relationships between hashtags must take all these issues under consideration.

In the rest of the paper, we present a metric for gauging the similarity of a pair of hashtags, and investigate the variations in this similarity over time. We apply our model on a real dataset and present our findings.

1.2 Dataset

We have access to the entire Twitter dataset for January 2011, used in [5]. It contains every tweet posted for that month, which is slightly more than 3 billion tweets. We will be using this dataset in all our experiments. Table 1 lists some statistics about the dataset. Our analysis will be restricted to tweets that contain at least one hashtag.

1.3 Initial findings

As an initial step, we parsed the JSON tweet data and extracted the counts of hashtag mentions in the entire dataset. Interestingly, the frequency distribution of hashtags follows a power law with a slope[2] of 1.2 (Figure 1). Figure 2 shows the distribution of the number of hashtags occurring in a tweet. Although a majority of tweets had zero hashtags (91%), we can see that this distribution is also roughly following a power law.

From the frequency distribution we can conclude that amongst the tweets that contain hashtags most are fleeting and not picked up by other users. However,there is a small fraction of hashtags that occur multiple times and are at the center of most relationships. Analysing these hashtags will provide the best idea about the structure of relationships between hashtags.

[2]We used LR to estimate alpha.

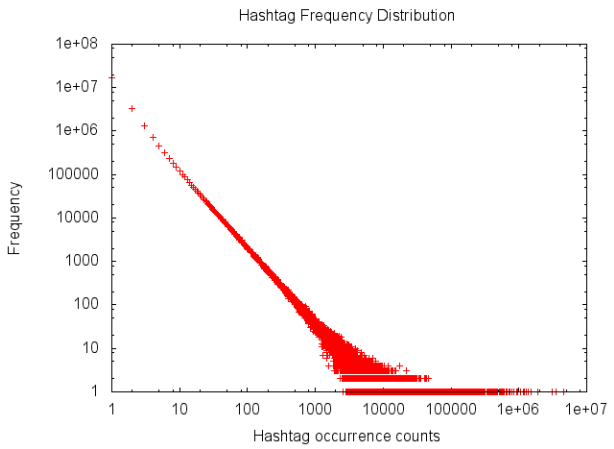

Figure 1: *Distribution of the occurrence counts of hashtags. It is clearly following a power law distribution.*

2 Prior Work

2.1 Topic sentiment analysis in Twitter

Wang, Xiaolong, et al [2] focus on hashtag level sentiment classification. The authors classify hashtags into three categories, namely topic hashtags, sentiment hashtags and sentiment-topic hashtags. In order to capture the relationship amongst hashtags they consider a graph with hashtags as nodes and create an undirected unweighted edge between two nodes if those particular hashtags appeared together in a single tweet.

The core extension we explore regarding [2] concerns the hashtag graph. The edges between the nodes of the graph are unweighted. This gives equal weightage to relationships between hashtags in a tweet that was re-tweeted thousands of times and a tweet that was not re-tweeted at all. Further, there is no mention of how the relationship between hashtags change with time. We seek to formulate a similarity metric between hashtags which encapsulates not only the co-occurrence frequency but also the signal from the common neighbors of the pair of hashtags.

2.2 Patterns of temporal variation in online media

Yang, Leskovec [3] have presented a study of temporal patterns associated with hashtags occurring on Twitter. They track how a particular hashtag's popularity grows and fades over time. They developed a K-Spectral Clustering algorithm which clusters the

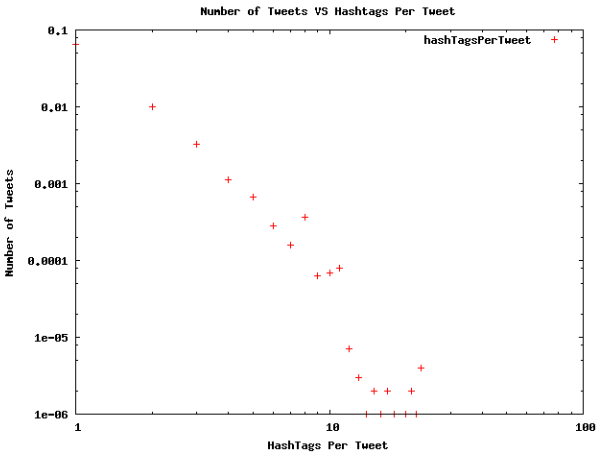

Figure 2: *Distribution of no. of hashtags per tweet (log-log scale). Note that it seems to be roughly following a power law.*

time series of various online content.

Although [3] presents a model for clustering the temporal patterns of occurrence counts of hashtags, it does not explore the temporal variation in the co-occurrence of pairs of hashtags. We explore this direction as the variation in the co-occurrence of hashtags can bring to light the strength of relationship between different topics.

2.3 Predicting the future with social media

Asur et al [4] utilize traffic from Twitter to predict the box office revenue of films and achieve better results than the Hollywood Stock Exchange, the industry gold standard. The authors study how buzz and attention is created for different movies and how their promotion strategies vary with time. They explore an application of using tweets and hashtags to gauge the society's reaction to a particular event and make predictions from it. We analyze the affinity sets of hashtags related to an event, as that would signal the "public engagement" generated by that event, and also give insights into what is being talked about in relation to that event.

3 The Hashtag Graph

A lot of research has been directed towards modeling users in a social network. Typically, these analyses take users as the nodes and edges between two users are created based on similarity in their usage of hashtags. This allows for the clustering of users based on their behavior, and can potentially be helpful in content recommendation.

However, in the context of analyzing the hashtags themselves, we consider a graph where the nodes are hashtags and edges are created based on some measure of similarity between those hashtags. The global structural properties of this graph like average degree and average clustering co-efficient reflect the overall nature of hashtag use, while local properties of a particular hashtag (degree, betweenness centrality, etc) depict the popularity and engagement generated by that hashtag.

A key question to consider is the criteria for creating an edge between two hashtags. [2] have simply constructed an edge between two hashtags if they co-occur in at least one tweet, thus constructing an undirected, unweighted graph. Initially, we sought to improve upon this model by considering the following scenarios to add an edge between two hashtags $\#h_1$ and $\#h_2$:

1. No. of times a user tweets $\#h_1$ and $\#h_2$ in same tweet. This is the traditional co-occurrence frequency for a pair of hashtags.

2. No. of times a user U_1 tweets $\#h_1$ and another user U_2 replies to that tweet with a tweet containing $\#h_2$. This signifies that $\#h_1$ and $\#h_2$ are frequently used in the same conversation between users.

3. No. of times a user tweets $\#h_1$, immediately followed by another tweet containing $\#h_2$. This signifies that $\#h_1$ and $\#h_2$ are frequently tweeted in the same context by a user. The intuition behind this signal is that users may choose to post about the same topic in multiple tweets.

However, we found that the simple co-occurrence measure itself produces many noisy connections between hashtags, and considering the scenarios 2 and 3 above leads to even more noise. Since the dataset contains all the tweets in a particular time span, it covers a diverse range of use of these hashtags, leading to many noisy co-occurrences. A key insight is that rather than considering complex scenarios for connecting two hashtags, it will be more fruitful to devise a method for ranking the edges formed by simple co-occurrence, and filter out the noisy connections. This has guided our approach to devising the similarity measure for pairs of hashtags (Section 4).

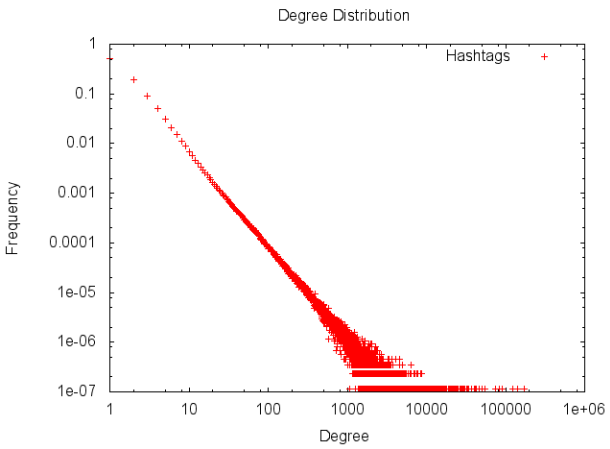

Figure 3: *Degree distribution of the hashtag graph created from the entire dataset.*

3.1 Preferential attachment in the hashtag graph

As a first step, we constructed an undirected, unweighted graph of all hashtags from the entire dataset, adding edges between pairs that co-occur in at least one tweet. The resulting graph has a degree distribution that closely follows a power law. This can be explained by a preferential attachment model of hashtag co-occurrence:

1. Suppose hashtags are created in the order $\#h_1$, $\#h_2$, $\#h_3$, ...

2. At the time of creating $\#h_j$, let d_i be degree of hashtag $\#h_i$, $i < j$.

3. Suppose $\#h_j$ is used together with m other existing hashtags.

4. Probability of a user linking a hashtag h_j with a previous hashtag h_i is proportional to the popularity of h_i, denoted by the number of times it has co-occurred previously with other hashtags, or d_i.

5. Thus, new hashtags are more likely to be used in conjunction with already popular hashtags.

This model of "rich get richer" explains why a small set hashtags become very popular. We will mainly focus on this long tail of popular hashtags in our subsequent analysis.

Hashtag 1	Hashtag 2	Co-occurrence count
ipad	iphone	212048
hiring	jobs	180209
egypt	jan25	169143
tcot	teaparty	121809
music	nowplaying	76562
Aquarius	ZodiacFacts	72660
Egypt	Mubarak	46952
Football	Highlights	45697
android	app	38025
car	sale	25558

Table 2: A selection of popular hashtag pairs

3.2 Popular hashtag pairs

Sorting the edges of the hashtag graph in descending order of co-occurrence counts of the hashtags gives us a picture of the popular hashtag pairs. Table 2 lists a selection of such hashtag pairs[3]. We can classify the pairs as:

- **Permanent associations between entities in the popular lexicon**: "hiring, jobs", "car, sale", "music, nowplaying" and "Aquarius, ZodiacFacts"

- **Associations formed out of popular culture references**: "ipad, iphone", "Football, Highlights" and "android, app"

- **Associations arising out of recent events**: "egypt, jan25", "Egypt, Mubarak" (the Jan 25, 2011 revolution in Egypt to overthrow President Mubarak) and "tcot, teaparty" (The Tea Party being a "top conversation on twitter" in Jan 2011)

We will examine a representative pair of each type in our experiments (Section 6).

The distribution of pairwise co-occurrence counts also follows a power law (Figure 4). This can be explained by noting that in addition to the preferential attachment model for creating a new pair of hashtags as described in the previous subsection, the user chooses between creating a new hashtag pair or retweeting an existing pair based on a preferential attachment model as well. Existing pairs will be retweeted with probability proportional to their

[3]We have manually selected the pairs in order to avoid the hashtags of objectionable/adult content that formed a majority of the top co-occurrence pairs.

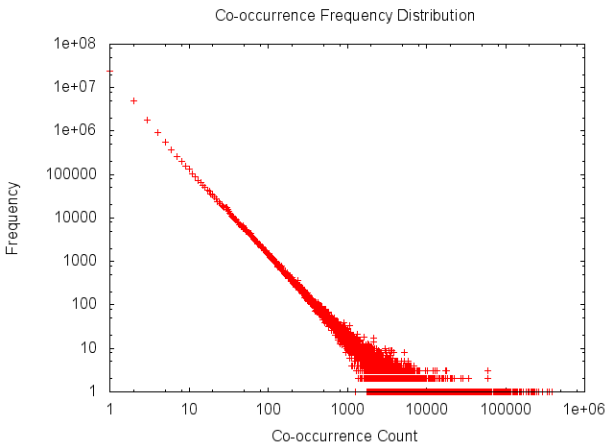

Figure 4: *Distribution of the co-occurrence counts of hashtag pairs over the entire dataset.*

current retweet count, hence the "rich get richer" phenomenon is at work over here as well.

3.3 Global properties of daily hashtag graphs

Next, we constructed a hashtag graph for tweets occurring on each particular day. Since our dataset spans the month of January 2011, we constructed 31 graphs (See Implementation details, Section 6.1). Table 3 lists some global properties of the hashtag graph for January 15, 2011. The graph is highly clustered, which can be inferred from the high average clustering co-efficient. It is also highly connected, as the largest connected component contains almost 80% of the nodes. Interestingly, all of the 31 daily graphs showed very similar values of these properties.

4 Hashtag Similarity

Co-occurrence in a tweet is a weak measure of the strength of the relationship between hashtags, as revealed by our analysis in the previous section. Since the hashtag graph is extremely noisy, the goal is to filter out the spurious edges and give higher weights to edges that signify a meaningful relationship. We present a similarity measure between hashtags that considers the neighborhood structure of the pair in question in addition to the simple co-occurrence counts.

We first describe two scores that form an integral part of the similarity measures, and then present two types of similarity measures using these scores:

Nodes	584512
Edges	1621963
Avg Clustering Co-efficient	0.346443
Fraction of nodes in largest WCC	0.779895
Diameter (approximate)	12

Table 3: *Global characteristics of the hashtag graph for Jan 15, 2011.*

4.1 Term weight

A Hashtag which occurs frequently in many contexts does not provide as much information about similarity with another hashtag as compared to hashtags that occur less frequently. The Twitter dataset has many hashtags that constitute "spam" which occur frequently but do not add any information to the strength of relations between hashtags. Weighting the hashtags in inverse proportion to their frequency should eliminate such spurious signals.

If the maximum frequency of any hashtag in the dataset is C, then we define the term weight as:

$$TermWeight(h) = log\left(\frac{|C|}{count(h)}\right)$$

where $count(h)$ is frequency of occurrence of the hashtag h in the entire dataset.

4.2 Pairwise co-occurrence

Intuitively, if a pair occurs with higher frequency, then we can assign higher confidence in the strength of the relationship between the two hashtags:

$$PairSim(h_1, h_2) = log\,(1 + count(h_1, h_2))$$

where $count(h_1, h_2)$ is the pairwise co-occurrence count of h_1 and h_2.

4.3 Direct Similarity

This similarity score captures the signal from the direct edge between two hashtags. If $\#h_1$ and $\#h_2$ have an edge between them, then the edge is more important if h_1 and $\#h_2$ occur are not very frequent. We define the direct similarity as:

$$DirectSim(h_1, h_2) = PS(h_1, h_2) * (TW(h_1) + TW(h_2)) \quad (1)$$

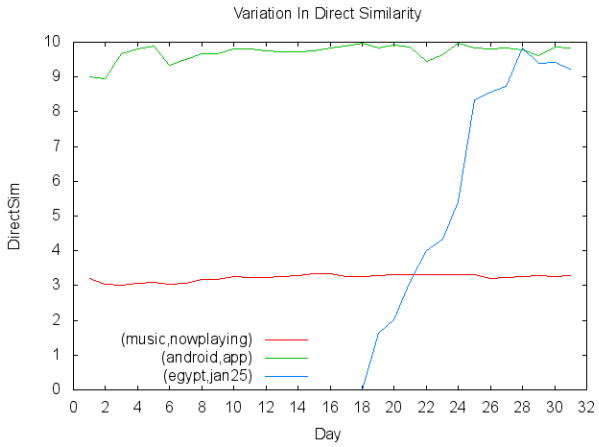

Figure 5: *Temporal variation in DirectSim values for three representative hashtag pairs.*

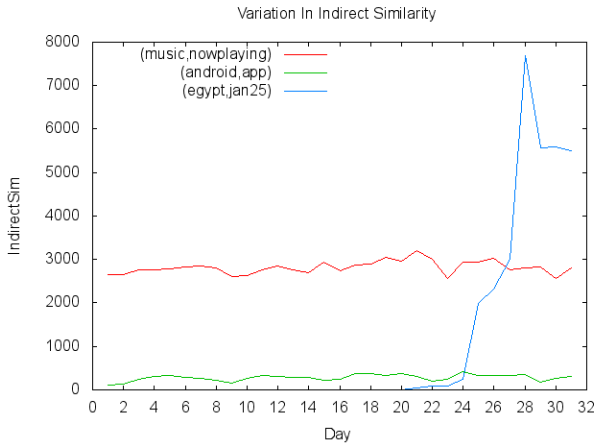

Figure 6: *Temporal variation in IndirectSim values for three representative hashtag pairs.*

Thus, the direct similarity is the pairwise similarity of the edge weighted by the sum of the term weights of the nodes.

4.4 Indirect Similarity

In addition to the direct relationship between a pair of hashtags, there is significant signal in the set of common neighbors of the two hashtags. We define the second similarity measure to encapsulate the signal from the common neighbors in the hashtag graph. Common neighbors with higher TermWeight provide more confidence to the strength of the hashtag pair.

Also, if the pairwise similarity of the neighbor with each of the two hashtags is similar, that is a higher signal of strength of that pair, as compared to when the pairwise similarities are highly skewed towards either one of the hashtag. If a hashtag n is a common neighbour between h_1 and h_2, then $PairSim(h_1, n)$ being close to $PairSim(h_2, n)$ implies a stronger correlation between h_1 and h_2, as compared to when the two pairwise similarities are very different.

$$IndirectSim(h_1, h_2) = \sum_{n \in nbr(h_1) \cap nbr(h_2)} \left(TW(n) * \frac{S(h_1, h_2, n)}{T(h_1, h_2, n)} \right) \quad (2)$$

where

$$S(h_1, h_2, n) = 1 + min(PS(h_1, n), PS(h_2, n)) \quad (3)$$

$$T(h_1, h_2, n) = 1 + max(PS(h_1, n), PS(h_2, n)) \quad (4)$$

We present the results of applying these similarity scores on the dataset in Section 6.

5 Temporal Analysis

The hashtag graph is dynamic: everyday, new hashtags become popular while old ones go out of use, depending on what is trending on that day. We can expect the similarity measures of pairs of hashtags to vary drastically with time. Trending hashtags allow us to see the popular topics on twitter on that particular day. However, it does not tell us anything about the relationship between hashtags. Analyzing such variations in the affinity sets can give insights about which pairs of hashtags are strongly correlated over long periods of time, and which pairs show only a temporary, "momentous" correlation, which could be explained by some related event, news outbreak, or even a marketing campaign.

We consider the temporal variation in similarity of the hashtag pairs by computing the scores on graphs formed over subsets of the dataset. We have chosen to divide the dataset according to days, so that all the tweets corresponding to a single day fall into one subset. Alternative schemes of dividing the dataset, for example variable number of hours or sliding windows [3], could be employed as well (See Future Work, Section 7.1).

The scores described in Section 4 are calculated for hashtag pairs occurring in each graph. This provides a set of DirectSim and IndirectSim scores for each pair of hashtags. The TermWeights of hashtags are computed based on their frequency in the com-

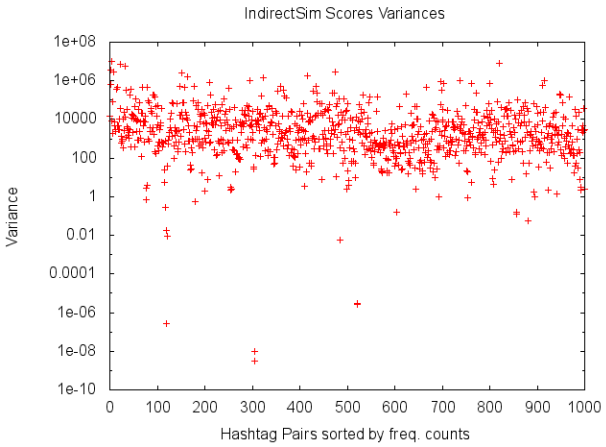

Figure 7: *Variances in the indirect similarity scores of the top 1000 hashtag pairs, sorted by frequency count.*

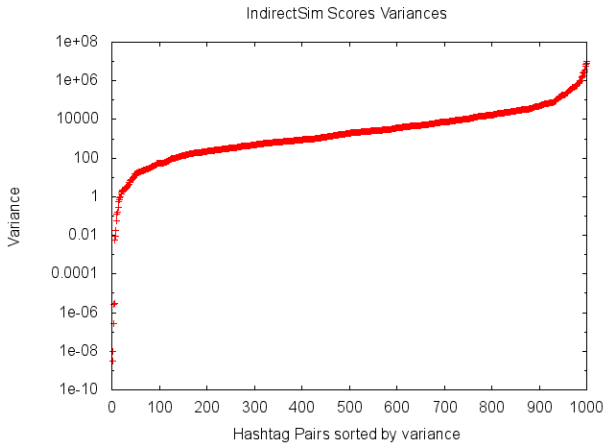

Figure 8: *Variances in the indirect similarity scores of the same 1000 edges as in Figure 7, but sorted in increasing order of variance.*

plete dataset, while PairSim values are based on the frequency of that pair in the particular time interval.

For a graph at time interval k, the direct similarity DS_k is:

$$DS_k(h_1, h_2) = PS_k(h_1, h_2) * (TW(h_1) + TW(h_2)) \quad (5)$$

and the indirect similarity IS_k is:

$$IS_k(h_1, h_2) = \sum_{n \in nbr(h_1) \cap nbr(h_2)} \left(TW(n) * \frac{S_k(h_1, h_2, n)}{T_k(h_1, h_2, n)} \right) \quad (6)$$

For a pair h_1, h_2, the sequence of scores

$$DS_1(h_1, h_2), DS_2(h_1, h_2), ...DS_K(h_1, h_2) \quad (7)$$
$$IS_1(h_1, h_2), IS_2(h_1, h_2), ...IS_K(h_1, h_2) \quad (8)$$

denote the temporal variation in the similarity strength. The variance of these sequences signifies the "stability" of the strength of the relationship between the h_1 and h_2. We use these statistics to evaluate our similarity model.

6 Experiments

In this section we present our findings from applying our model for temporal variations of hashtag similarity on the Twitter dataset.

6.1 Experimental setup

Due to the large size of the dataset, we had to write scripts to parse it using a multi-threaded approach. We worked on 64 core, 1TB RAM machines in the Stanford InfoLab for parsing the dataset and computing similarity scores.

We made heavy use of the efficient graph management library SNAP[4] to perform the graph computations. A significant part of our project involved making extensions to the SNAP platform to enable analysis on sequence of graphs in a parallel fashion. We were able to efficiently load the entire dataset into memory, divide it into chunks for temporal analysis, and run our algorithms on each chunk in parallel. This allowed us to easily scale the algorithms to such a large dataset.

6.2 Temporal variation in similarity

We computed the direct and indirect similarity scores for every edge of all the graphs generated by the temporal paritioning of the dataset. Figures 5 and 6 plot the variation in the direct and indirect scores, respectively, of three hashtag pairs that are representatives of the different classes described in Section 3.2.

We can see that the pairs "android, app" and "music, nowplaying" hav an almost constant direct and indirect similarity, since these pairs have a relation strength that is independent of time. On the other hand, the pair "egypt, jan25" starts with

[4]One of the authors is on the SNAP development team

Hashtag 1	Hashtag 2	Variance
JobHunting	hiring	40.03
Football	Highlights	41.20
Careers	tweetmyjobs	57.38
financial	services	79.47
Cancer	ZodiacFacts	121.60

Table 4: Hashtag pairs with low variance

Hashtag 1	Hashtag 2	Variance
egypt	jan25	6966675.71
Mubarak	jan25	1013495.92
tcot	teaparty	190212.79
Bears	Packers	1405247.90
Verizon	iPhone	183689.09

Table 5: Hashtag pairs with high variance

zero similarity, but shoots up to high values around Jan 25th, closely following the political revolution in Egypt. Interestingly, "android, app" has a higher direct similarity but lower indirect similarity than "music, nowplaying". This implies that "android" "app" co-occur in the same tweet more frequently, while "music" and "nowplaying" have more common neighbors which co-occur with both of these tags. Finally, the indirect similarity value of "egypt, jan25" rises to a very high value in the days following Jan 25th, which can be explained by these hashtags being mentioned in many common contexts on Twitter as a consequence of the revolution.

6.3 Evaluation

To evaluate our similarity metric, we consider the indirect similarity scores of the top 1000 most frequent hashtag pairs over each time interval. The variance of the values described in equation (8) for each such hashtag pair denotes the stability of the similarity model. First, we see in Figure 7 that there is no correlation between the frequency count and the variation in indirect similarity of the hashtag pairs. On the other hand, Figure 8 shows that a majority of the variances are in a particular range, with a few outliers on both ends.

Table 4 lists some of the hashtag pairs which are outliers on the lower end, while Table 5 lists out some pairs that are outliers on the higher end. We can see that the pairs of hashtags exhibiting low variance are the ones expected to have a strong relationship irrespective of time, while the ones exhibiting high variance are ones which depict momentary connections based on some recent event. Clearly, we can use this similarity model to track the quality of engagement generated by a given hashtag. Specifically, we can track which co-occurring hashtags have a strong similarity score with the given hashtag, while which ones are caused by momentary events.

7 Conclusion

We have presented an investigation of the characteristics of the Twitter hashtag graph which have been largely unexplored, to the best of our knowledge. Although the hashtag graph is extremely noisy, there is significant information in the pairwise co-occurrence of hashtags, which has a direct mapping to associations between entities of popular interest. The quality of a similarity metric is proportional to its capacity to eliminate the noise and uncover hashtag pairs that exhibit meaningful relationships. Such a tool for extracting insights from hashtag data would be an invaluable addition to knowledge bases that power many semantically aware systems.

7.1 Future Work

We have considered data spanning only one month. Clearly, we can gain further insights into the quality of the similarity metric by performing the analysis over a larger time span. Additionally, we have considered only disjoint time intervals; it will be interesting to see the results arising out of time intervals taken as "sliding windows" over the time period in consideration.

Our evaluation metric was based on handpicking a selected set of hashtag pairs and analyzing the quality of the similarity score on those pairs. However, the hashtag dataset is extremely noisy and contains many objectionable/spurious tags. We did not preprocess the dataset to remove such content. A more airtight evaluation scheme would either take into consideration these spurious tags, or do away with them at the preprocessing stage itself.

References

[1] Fabian Suchanek and Gerhard Weikum. 2013. "Knowledge harvesting in the big-data era." In Proceedings of the 2013 ACM SIGMOD Interna-

tional Conference on Management of Data (SIGMOD '13).

[2] Wang, Xiaolong, et al. "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[3] Jaewon Yang and Jure Leskovec. 2011. "Patterns of temporal variation in online media." In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 177-186.

[4] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

[5] Myers, Seth A., and Jure Leskovec. "Clash of the contagions: Cooperation and competition in information diffusion." Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012.