ANALYSIS OF TWITTER HASHTAGS' GEOGRAPHIC PROPAGATION

A Thesis

by

ASHRAF AHMED IBRAHIM ABDELHALIM

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE

Chair of Committee,	James Caverlee
Committee Members,	Richard Furuta
	Eduardo Gildin
Head of Department,	Dilma Da Silva

December 2015

Major Subject: Computer Science

ABSTRACT

The goal of this work is to study the geographic propagation patterns of Twitters' hashtags. In order to analyze the hashtags' diffusion patterns, we look at the globe as a graph consists of a large grid of locations and use two different approaches to study the hashtags' behaviour. The first approach is to consider the locations on the global grid as variables (or features) and the individual hashtags as the examples. This viewpoint of our dataset allows us to perform dimensionality reduction techniques to reduce the size of the dataset without much loss of information and to identify the more influential locations. The second methodology is to transform the global grid into a undirected weighted graph and compute the influence curves associated with the hashtags propagation and their properties. We show that the influence curves of different classes of hashtags have similar patterns. In addition, we show that the influence curve can be approximated adequately using only six Chebychev polynomials.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. James Caverlee for his help and support during the time I have been student at the computer science and engineering department at Texas A&M University. I am highly appreciative for his patience, guidance and knowledge. It has been a privilege to part of the information retrieval laboratory and working with Dr. Caverlee and the rest of the lab. members.

In addition, I would like to thank Dr. Eduardo Gildin, from petroleum engineering department, for his generous support and for giving me the opportunity to collaborate with him and his students in very interesting projects. I would like also to thank Dr. Richard Furuta for his help and support and for being part of this research. Special thanks to my former colleague in the info. lab. Krishna for introducing me to this work.

Finally, special thanks to Dr. Nimir Elbashir, from Texas A&M Qatar and Dr. Mhamoud Elhalwagi, from the Chemical Engineering department for their support and for the opportunity to work with them and apply some of the things I learned in the info. lab. in the field of chemical engineering.

NOMENCLATURE

$D(G)$	Density of graph G
$F(p)$	The persistence parameter
$f(r)$	The frequency of data distribution of rank r
$M(p)$	Stickiness parameter G
$p(k)$	The influence curve
T_n	Chebychev polynomial of degree n
w_{ij}	A weight of a graph edge

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

Social media has become an important platform that allows users to communicate their ideas in addition of being a source of news and many other social activities. The growing importance of social media in our daily lives has attracted a lot of attention of researchers in fields such as computer science, finance and marketing, sociology, among many others. Understanding the spread of ideas and memes in a social network is important for a range of significant activities, including understanding the reach and impact of news, the potential impact of commercial marketing, and deeper sociological issues of how people connect, self-organize, and communicate.

In an exciting new direction, we now have access to fine-grained data that encodes not only when a meme has been adopted, but also where that meme has been adopted. In particular, this thesis opportunistically mines the worldwide dissemination of Twitter hashtags, as a first step toward building robust models of spatio-temporal meme diffusion. A Twitter hashtag is an annotation associated with a message that is generated by the user posting the message. Typically hashtags are a single word or a simple combination of words prefixed by the symbol #. Many users use hashtags as a way to categorize messages posted on Twitter and to promote topics or ideas. Recently there has been a growing interest in the dynamics of hashtag propagation in social networks, e.g., [1, 4, 6, 11, 13]. And in this work, we continue this investigation by analyzing the geographic impact of hashtag propagation; since Twitter provides a GPS-enabled tagging service which is used in around 1% all tweets, we can study how geographic factors impact propagation.

1.1 Research summary

The goal of this work is to investigate the spatial dynamics of Twitter hashtags and their propagation patterns through a geographic social network. In this thesis, we are interested in studying the geographic propagation of hashtags by looking at the world as a large grid of cells of equal area, where each cell represents a geographic location. Over this grid, we ask questions like:

- How informative is the data and how much redundancy is there in the dataset? How much correlation is between locations? How to detect anomalies in the data?
- Do different kinds of information spread differently online? What do these "influence curves" look like and how "sticky" or "persistent" are different memes?
- Can we model and study these properties over a geo social network? And can we approximate these influence curves?

We address these questions through two parallel approaches:

Approach 1: Analysis using Dimensionality Reduction. Firstly, for each hashtag, we consider the locations as its attributes (or features) and the value of an attribute as the number of times the hashtag appears in the location. Then we compute principal component analysis (PCA) [15] of the normalized and the scaled data in order to identify the most significant locations and study the distributions and the correlations between locations. Principal component analysis allows us to eliminate the features with small loading values and retain only a small number of features (or locations in this case) of more important role in the data. This is, in turn, very useful in reducing the original dataset to a much smaller subset without

graph	num. of nodes	density	max. degree	ave. degree	spec. radius
graph 1	6135	0.004303	1055	26.393	155.817
graph 2	5140	0.00362	732	18.6147	106.774
graph 3	4509	0.003411	564	15.377	85.873
graph 4	4149	0.003196	465	13.2552	73.032

Table 1.1: The structure of the four graphs

much loss in the total variance and therefore we can perform statistical analysis on more manageable dataset.

Approach 2: Study of the Influence Curve. Secondly, we consider the cells of the global grid as nodes of undirected graph and for any two nodes (or locations) we count the number of pairs of users that mentioned each other in their tweets at least once. If the number of such pairs exceeds a certain threshold, we consider an edge between the two nodes and therefore for a different threshold we have a different graph. Table 1.1 shows the basic properties of the four graphs used in this thesis. Then, we compute the influence curve [11] of the hashtags which gives the probability of a hashtag to appear in a location after appearing in a number of the location's neighbors. We show that the influence curve can be approximated by a fifth degree polynomial, in other words, we need only six numbers to fully describe the curve.

1.2 Related work

Twitter is a popular online social networking and microblogging service and there has been considerable research using Twitter as platform to study and model the information diffusion processes over social networks [14, 7]. One of the interesting problems is to study the variations of how different memes and ideas flow in online social community. Romero et. al. [11] have studied the notions of “stickiness” and "persistence" and how they vary with respect to the class of the hashtag. Using a language-based approach, Cunha et. al. [1] studied the propagation of hashtags in Twitter and showed quantitative similarities between online and offline communities. Kamath et. al. [6] studied the adoption of hashtags' based on time, distance and location. Furthermore, they examined the focus, entropy, and spread of the hashtags' spatial propagation and showed that most hashtags spread over small geographical areas but at high speeds. Kamath et. al. [4] proposed a model to predict the popularity of hashtag in a given location using a reinforcement learning approach.

In order to understand the spread of social media, Kamath et. al., [5] attempted to model this global phenomena by studying the geo-spatial properties of a dataset of geo-tagged tweets . They showed that distance is important in hashtag adoption and that hashtags are in general a local phenomena. Tsur and Rappoport [13] proposed a hybrid linear regression method in order to predict the spread of a hashtag within a given time period and they used temporal and topological features to achieve good performance. In order to characterize Twitter hashtag cascades based on different topics, Rattanaritnont et. al., [10] discussed distributions of user influence, cascade ratio and tweet ratio. In addition to that, they used K-means clustering based on tweet ratio and cascade ratio and showed that there are three different patterns for the topics selected.

2. DATA DESCRIPTION AND SETUP

We have collected Twitter data for two months of September and October of 2011 and considered the geo-tagged tweets only. For the geo-tagged tweets, each one is tagged with geographic information of a latitude and longitude indicating the location of the user where she posted the tweet. This geographic information takes the form <`hashtag, time, latitude, longitude`>. The geographic coordinate system we use in this work is Universal Transverse Mercator (UTM) [8], which is a two-dimensional system that maps any location in the world to a Cartesian coordinate. The UTM system is used because it identifies locations regardless of the vertical position and it only considers the horizontal position.

We call this structure *the global grid* and each cell in this grid is an actual location in the world map. The size of cell (or location) in the grid is about 10^8 square meter and we considered all locations with at least one appearance of a hashtag among the group of the hashtags used in this thesis. This work focuses on a collection of the most frequently used hashtags and we consider the top-400 hashtags. We are only considering hashtags that were "alive" during the two-month period even though the lifespan of most of these hashtags extends beyond this time period. The total number of locations in the dataset is 9,162 locations around the globe; the top-100 locations with respect to hashtag frequency are shown in Figure 3.5(a). In addition, we classify the hashtags into four groups based on the topic. The groups (or classes) are: politics, entertainment, sports and technology.

The two main studies conducted in the work are:

- First, we consider the dataset containing the number of times each hashtag appears in a location. The dataset contains 9,612 variables of integral values and 400 samples. We study the data using PCA in order to reduce the high dimensionality and identify the most “influential” locations.

- Second, we build four different graph structures – where nodes are from the set of global grid cells and edges are based on the social interaction of users in those locations – and investigate the propagation patterns of hashtags. We compute the exposure curves of each each graph and their parameters and approximations.

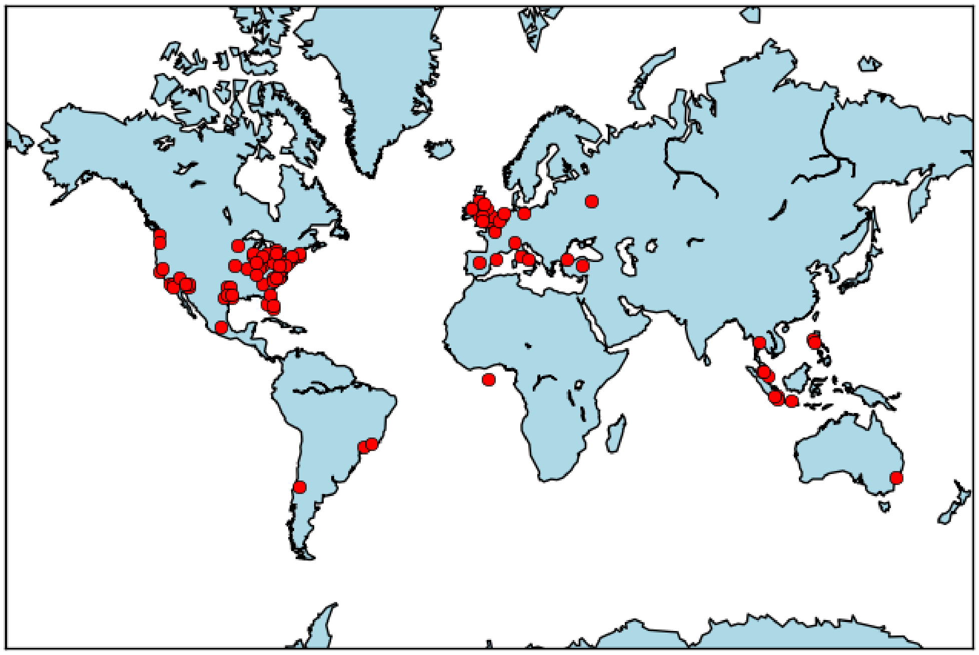


Figure 2.1: Top 100 locations based on the hashtag frequency. We notice that the location with high frequencies are clustered most in North America and Europe where a higher percentage of the population use the social media.

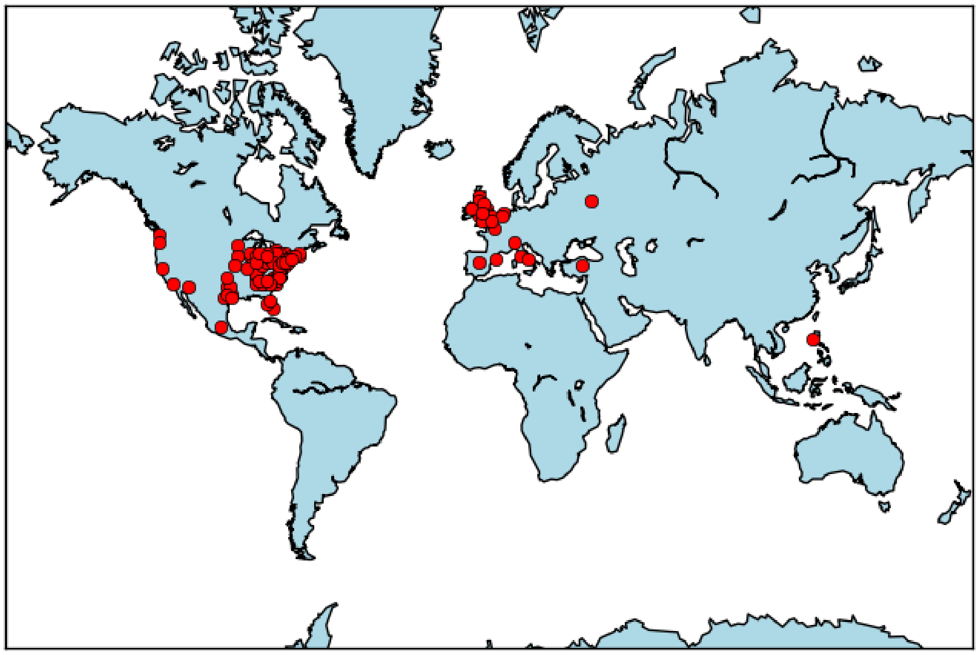


Figure 2.2: Top 100 locations based on the first component of PCA. We see that the distribution of the location around the globe is similar to their distribution based on frequency. But, on the other hand, far less number of locations are actually outside North America and Europe. This means that the locations in North America and Europe not only have high frequency but they are more "influential" as well.

3. DIMENSIONALITY REDUCTION

When the dimension of the data space is large, it is very difficult to interpret and visualize the data in addition to the problems that arise from the complexity of working with high-dimensional spaces. Therefore, in order to avoid the problem (or the curse) of dimensionality, we need to reduce the dimension of the space without much loss in the information in the data. One way to reduce the dimension is by keeping the most relevant variables from the original dataset (what is known as feature selection). Another method is by exploiting the redundancy in the data by finding a smaller set of variables each is a linear combinations of the original variables. The new set of variables contains almost the same information as that of the original input variable. Note that, the principal component analysis we use in this work implements the second methodology.

3.1 Principal component analysis

Principal component analysis (PCA) is a very popular statistical procedure for dimensionality reduction [3]. PCA is the transformation of a set of observations (examples) of possible correlated random variables into another set of uncorrelated variables. The new set of variables is called the principal components and they are linear combinations of the original set of variables. We are interested in using PCA to reduce the number of independent variables in order to identify the most influential locations.

The main idea of principal component analysis is to maximize the variance of a linear combination of the variables. Therefore, the first principal component is the linear combination with maximal variance. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on.

If we let $x_1, x_2, \ldots, x_n$ denote the independent variables, $x_i \in \mathbf{R}^m$, which are in this case the locations in the global grid, then we need to find a new set of variables $y_1, y_2, \ldots y_k$, $k << n$ which are linear combinations of x_i's without much loss of information in the original dataset. In other words, we are interested in finding $y_1, y_2, \ldots y_k$ such that

$$y_i = w_{i1}x_1 + w_{i2}x_2 + w_{i3}x_3 + \cdots + w_{in}x_n \tag{3.1}$$

such that the total variance $\sum_i \mathrm{var}(y_i)$ is maximized. The variables y_i's are the principal components and w_{ij}'s are the coefficients (or loadings). Let $X \in \mathbf{R}^{m\times n}$ be the data matrix, then the first principal component y_1 is given by $y_1 = Xw_1$ where w_1 is the solution to the following optimization problem

$$w_1 = \mathrm{argmax}\{wX^tXw : ||w|| \leq 1\} \tag{3.2}$$

where $||.||$ is a suitable norm to measure the variance. The second loading vector w_2 is given by solving the same optimization problem above and replacing the matrix X by

$$X_2 = X - Xw_1w_1^t.$$

Continuing in this manner, we can find all the loading vectors and hence all the principal components needed to reduce the dimension of the data.

PCA is closely associated with singular value decomposition (SVD) [2] where the singular values are the eigenvalues of the matrix X^tX. SVD currently is the standard way of computing PCA unless the number of required components is small. In addition to dimensionality reduction and data visualization, PCA can be used for clustering and outlier detection among many other applications.

3.2 Reducing location data dimension

The data generated from social media is usually very large and complex that the traditional tools for data analysis and processing are not suitable. In order to perform any statistical analysis or visualization of the data, both the volume of the data and the dimension of the space should be of manageable size. Dimensionality reduction is a fundamental technique to reduce not only the dimension of the data space but also the size of the dataset with minimal loss of information. Therefore, we need to apply dimensionality reduction methods to our dataset in an attempt to resolve the following questions:

- Since the volume of the data is large, is it possible to reduce the size of the data without much loss of information in the data? (for example, to reduce storage space needed).
- Is it possible to reduce the data dimension to improve data visualization?
- Can we remove redundancy and multi-collinearity in the data? (to improve machine learning models performance, for example).
- Can we detect outliers in the data?

In this part of the work, we consider a dataset that contains relatively small number of samples (around 400 samples) and very large number of features (around 9,612 variables). Each sample in the dataset represent a hashtag where the value of each variable in the sample is the number of times the hashtag appeared in the location. Therefore, the data matrix is sparse because any hashtag does not appear in the majority of the locations.

If we have a high-dimensional dataset and X is a random variable, the way we measure the information is by computing the total variance $\mathrm{var}(X)$, where

$$\mathrm{var}(X) = \frac{\sum(X - \bar{X})^2}{n-1}$$

and

$$\bar{X} = \frac{\sum X}{n}.$$

The basic idea of the principal component analysis is to find the direction that maximizes the total variance when the data is projected into that direction. This is usually done by considering the covariance matrix

$$\mathrm{covar}(X_i, X_j) = \frac{\sum(X_i - \bar{X}_i)(X_j - \bar{X}_j)}{n-1}.$$

When we apply principal component analysis (PCA) to our dataset generated from Twitter corpus we reduce the dimension from 9,612 to a much smaller number. When we apply PCA to the normalized data, we find that we need only around 100 principal components to retain more than 90% of the variance in the data, see Figure 3.2. On the other hand, as for the scaled data, we need around double that number to reach the 90% threshold of the total variance. Clearly from this observation, we conclude that there is a lot of redundancy in the dataset where we need only around 1% of the total number of variables to retain more than 90% of information in the data.

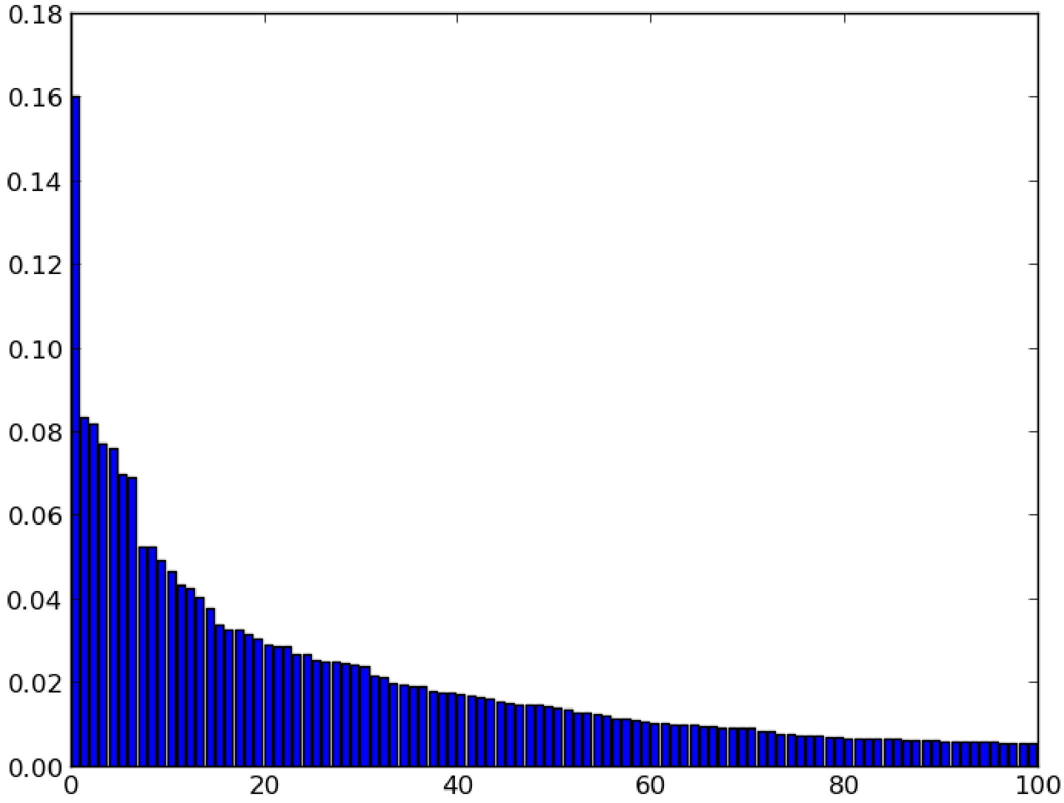


Figure 3.1: The variance of the data of the first 100 components from the normalized data. We see that the first component alone represents around 16% of the total variance and the first two components combined represent around 24% of the total variance. This usually implies that there is a lot of redundancy in the information the data has.

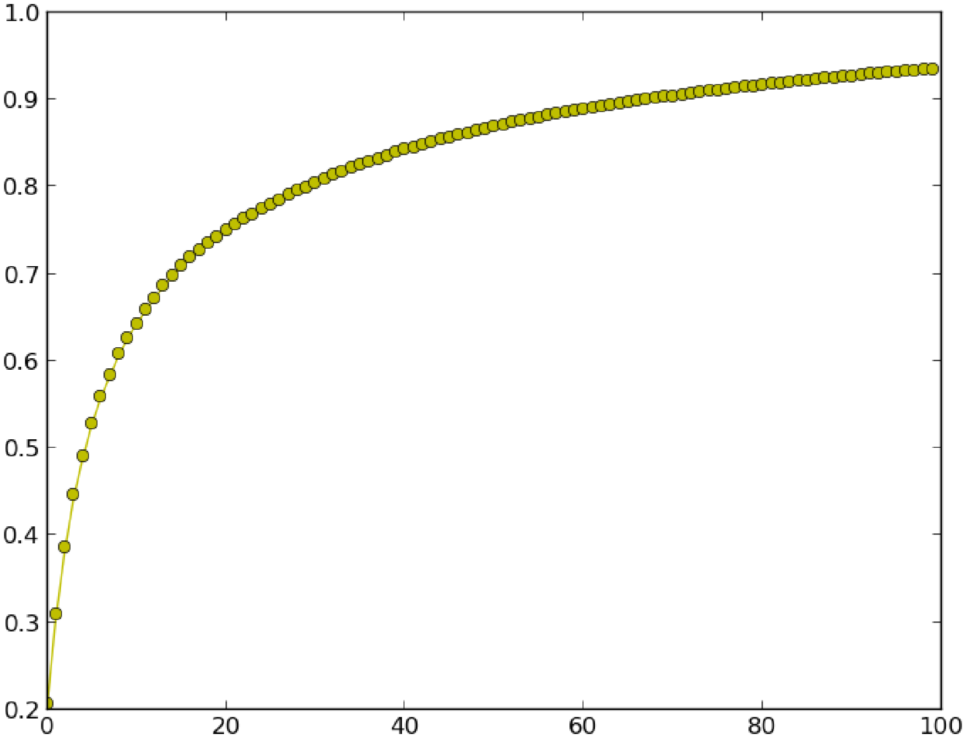


Figure 3.2: The accumulative total variance of the data of the first 100 components. We see that the first 20 principal components of the data represent more than 70% of the total variance which again validates the observation that there is high information redundancy.

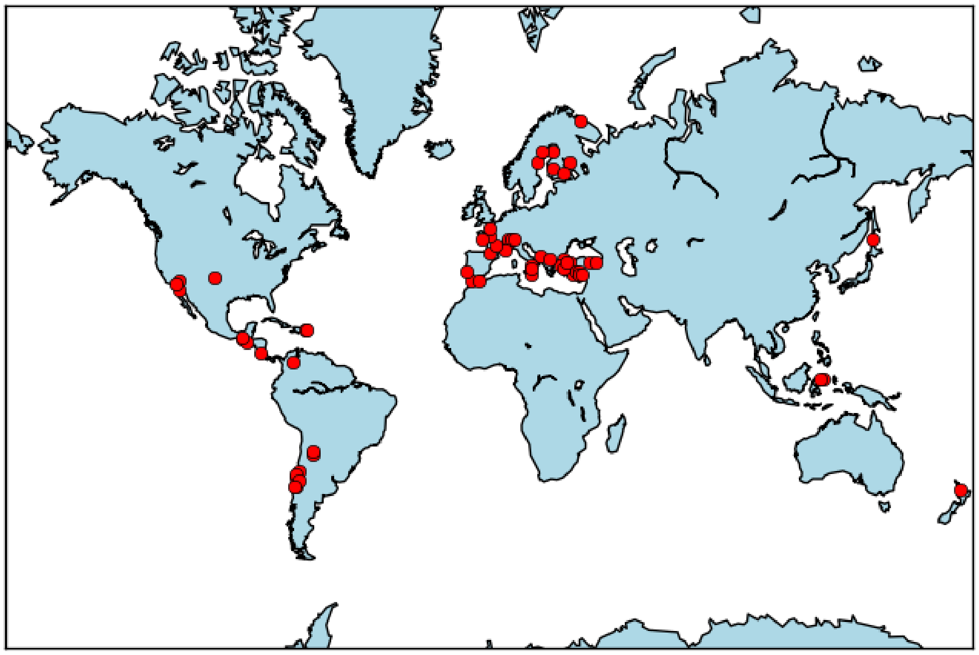


Figure 3.3: The top 60 locations of the first two principal components from the scaled data. The combined components give us a different distribution of the influential locations.

Using the first two principal components (which represent around 25% of the total variance) of the normalized data, the top-100 locations are distributed around the world similar to the distribution of location with high frequency of hashtag appearance. They are both concentrated in North America and Europe with 10% of locations in frequency group are outside North American and Europe whereas only 1% of locations in the PCA group are outside the two continent, see Figures 2.1, 2.2. If we only consider the data from the first two components and look at the distribution of the most "influential" locations, the distribution, in this case, is quite different as it is shown in Figure 3.3. We see in Figure 3.3 more locations appear in southern Europe and South America and far less locations in North America.

Figure 3.4 shows the scatter plot of the loadings of the first two principal components. We notice that the values of the majority of the loadings are clustered around 0. In addition to that, we see few values which are relatively high in both components.

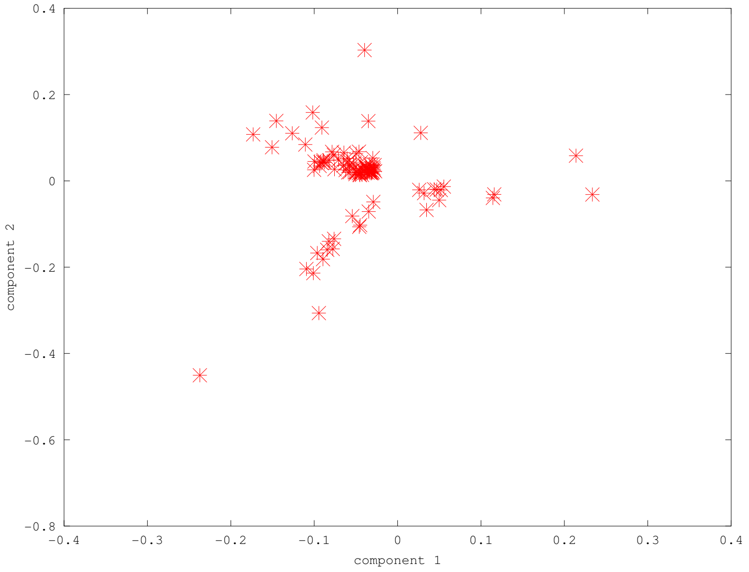
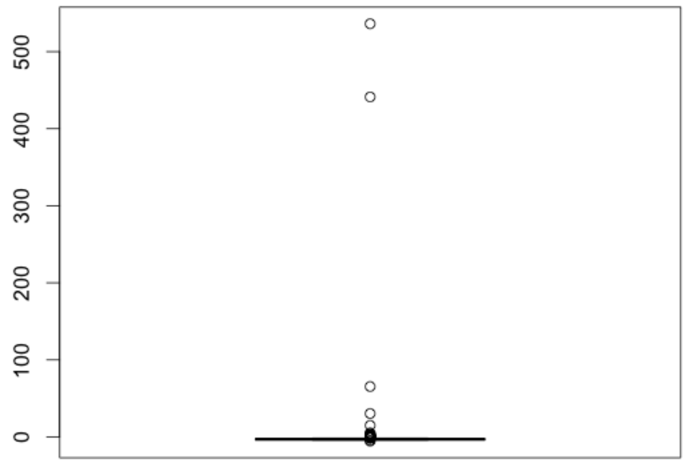
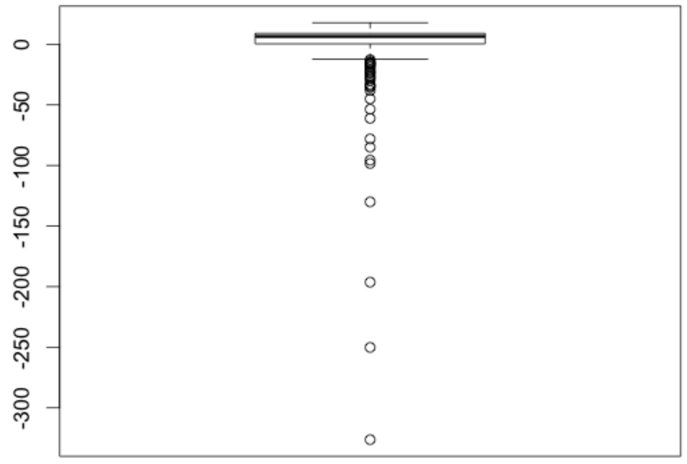


Figure 3.4: The scatter graph of the loadings of the first two components of PCA.

Figure 3.5 is the box plot of the loadings values of the first two principal components of the scaled dataset. As in the scatter plot, Figure 3.4, we again notice that most of the loadings values are clustered around 0. The box plot allows us to get an overall picture of how the data is distributed and its relation to some fundamental statistics like the mean and median. Furthermore, the box plot allows to identify outlier and other anomalies. We see a couple of outliers in the box plot (a) of the first component and similarly for the box plot (b) of the second component. Note that the majority of the loading values of the first component are positive numbers where, on the other hand, the loading values of the second principal component are mostly negative numbers. But in both cases, the majority of the absolute values of the loadings are less than 100 with the exception of few outliers. We conclude that, in the first component, most of the variables have mostly positive effect and for the second component the effect is mostly negative.



(a) Box plot of the first component.



(b) Box plot of the second component.

Figure 3.5: Figure (a) and Figure (b) show the box plot of the loadings of the first and second components. We see that most of the values are clustered around 0 with exception of few outliers.

4. THE INFLUENCE CURVE AND ITS APPROXIMATION

The general question we would like to investigate is: do different kinds of information spread differently online? The problem is that it has historically been difficult to evaluate the answer to this question quantitatively. In this thesis, we are interested in how, in a social network, the physical distance affects the way information propagates. First, we discuss the relationship between geographic locations and quantify such a relationship based on their hashtags' adoption. Second, we look at how the information spreads in different type of networks based on geographic location and how to quantify these spread patterns. In this thesis, we look at what is known as the influence (or exposure) curve which represents the probability of a location "adopting" the information after being "exposed" to it a number of times.

4.1 Relationship between locations

In this section we discuss the relationship between locations in terms of hashtag adoption. We are considering two methods to quantify such relationship between different locations. The first is based on the notion of Jaccard similarity which is the fraction of hashtags shared in the locations to the total number of hashtags in the locations. The second approach is based on the adoption time lag occur between locations.

Using the first method, given two locations, we look at the commonality of hashtags adopted in those two locations. The natural question we are interested in is to what extent the distance between locations affects the hashtag similarity. We define the hashtag similarity as follows

Definition 4.1.1. *If H_{l_1}, H_{l_2} denote the set of unique hashtags adopted in locations*

l_1 and l_2 respectively, then

$$\textit{Hashtag Similarty} = \frac{|H_{l_1} \cap H_{l_2}|}{|H_{l_1} \cup H_{l_2}|}.$$

Clearly, the hashtag similarity is a value between 0 and 1 and the higher the number, the stronger the similarity between locations. The relationship between hashtag similarity and the distance between locations is shown in Figure 4.1.

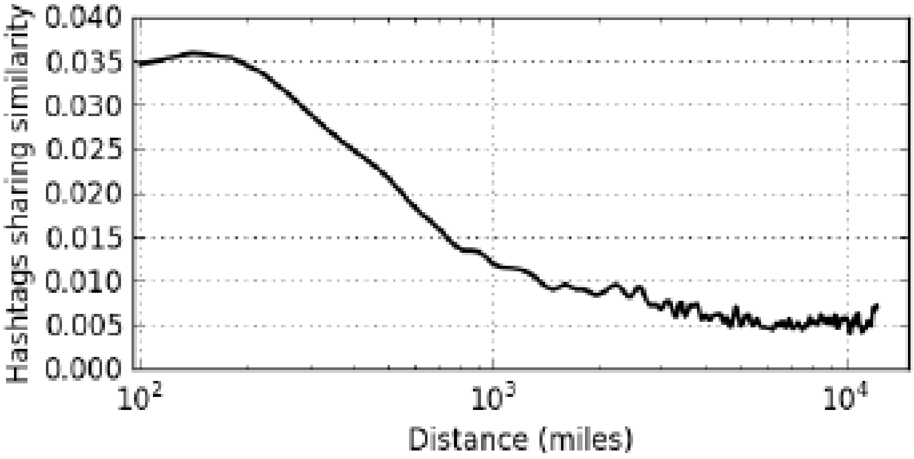


Figure 4.1: A plot of hashtag similarity with respect to distance between locations.

From Figure 4.1, we observe a strong correlation between hashtag similarity and the distance where the closer the locations are, the more likely the hashtag is adopted in the locations. We also see that the similarity decreases significantly as the distance increases and after certain distance, the increase in the location distance has less effect on the similarity. This observation is more intuitive as the location closer to each other are likely to share similar language, culture and common interests.

While two locations that are close to each other are more likely to adopt the same hashtag, then the next question is how soon those two locations will adopt the hashtag. We define the adoption lag in terms of the difference in time the hashtag appeared in the locations. We are trying to use the adoption lag notion to measure the temporal similarity of locations.

Definition 4.1.2. *Let $t^h_{l_1}, t^h_{l_2}$ be the first time hashtag h is adopted in locations l_1 and l_2 repectively, then*

$$Adoption\ Lag = \frac{1}{|H_{l_1} \cap H_{l_2}|} \sum_{h \in H_{l_1} \cap H_{l_2}} |t^h_{l_1} - t^h_{l_2}|.$$

The above definition measures the total time differences of hashtag adoption in the two locations normalized by the number of the common hashtags. The lower the value of hashtag adoption is, the similar the locations are and hence the common hashtag appears in the location almost at the same time.

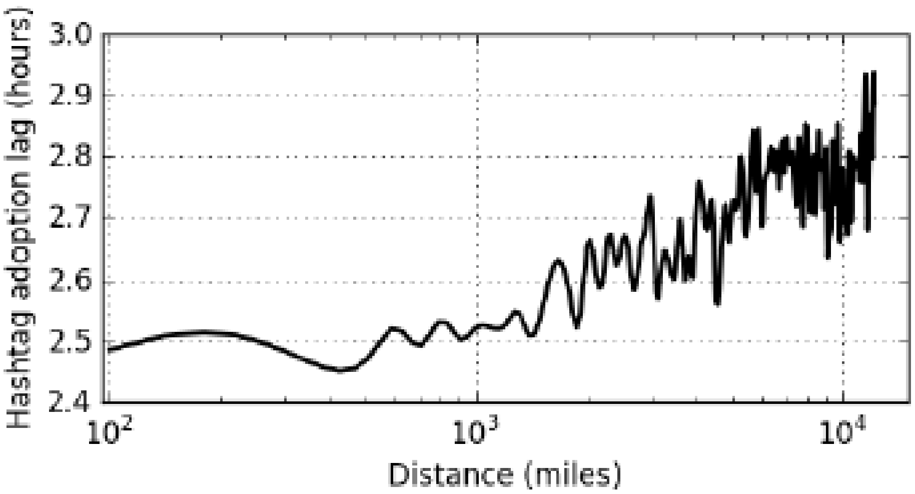


Figure 4.2: A plot of adoption lag with respect to distance between locations.

We see in Figure 4.2, for shorter distances (less than 500 miles), there is very little change in the adoption lag which indicates that the locations close to each other are likely to adopt the same hashtag almost the same time. On the other hand, for further distances (more than 1000 miles), we have a positive correlation where overall the adoption lag increase with the distance.

Based on these observations, we build several graphs based on the distance between locations in terms of how connected they are on Twitter. As background, the following section highlights some graph theory fundamentals.

4.2 Graph theory

A graph is a mathematical model that is used to represent pairwise relationship between objects (called nodes or vertices) and this relationship is represented by an edge. Graph theory has been applied to many fields in mathematic, science and technology including electrical engineering (in communication networks and coding theory), computer science (in algorithms and computability theory) and operations research (in scheduling). In information retrieval, graph theory is a fundamental tool to study many problems in social networks such as clustering, community detection and visualization of networks.

Formally, a graph G is an ordered pair $G = (V, E)$ where V is the set of vertices (or nodes) and E is the set of edges (or arcs or lines). If an edge $e \in E$ connecting the nodes $u, v \in V$, we write $e = (u, v)$, hence $E \subset V \times V$. We say a graph $G_1 = (V_1, E_1)$ is a subgraph of G if $V_1 \subset V$ and $E_1 \subset E$ (assuming $E_1 \subset V_1 \times V_1$). A path in graph G is a sequence of edges $e_1, \dots, e_n$ where e_i, e_{i+1} share a vertex. A graph is connected if there is a path between any two vertices.

Definition 4.2.1. *A connected component of a graph G is a maximal connected subgraph of G (i.e. it is not a subgraph of another connected subgraph of G).*

A connected graph has only one connected component. The notion of connectivity of a graph is very important in network analysis as it is a way to measure the graph robustness as network. The strongest connected graph is the complete graph where every edge is connected to every other edge in the graph. The adjacency matrix A of a graph G is the matrix $A = [a_{ij}]$ where

$$a_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in E \\ 0 & \text{otherwise.} \end{cases}$$

The Laplacian L of the graph G is defined to be the matrix $L = D - A$ where

$$D = \operatorname{diag}(\deg(v_i))$$

and $v_i \in V$.

Definition 4.2.2. *The spectral radius $\rho(L)$ of the graph G is the eigenvalue of the Laplacian matrix L with the largest absolute value.*

The spectral radius is also another important measure of the robustness of the network as the smaller the spectral radius the more robust the network. In addition, the spectral radius plays a crucial role in modeling propagation in many graph, for example, in modeling virus propagation network. Another measure of robustness of the network is the graph density which defined below.

Definition 4.2.3. *The density $D(G)$ of a graph $G = (V, E)$ is defined as follows:*

$$D(G) = \frac{2|G|}{V(V-1)}.$$

The connectivity and robustness of the network are ways to asses the vulnerability of the network to random failure. Another criterion of measuring such vulnerability is the degree distribution of the graph. Ideally, the degree distribution of a graph needs to be normally distributed where the degrees of the vertices are fairly close to an average degree. Many real-wold networks such as social networks and gene regulation networks do not have normally distributed degrees and there is usually a small number of vertices with high degrees.

4.3 The graph structure

We divide the globe into square grids of equal area using Universal Transverse Mercator (UTM), a geographic coordinate system which uses a 2-dimensional Cartesian coordinate system to map locations on the surface of the globe. The nodes of the network graph are n locations around the world (a bi-directed weighted graph). We say there is an edge from location i to location j if there are k users in location

i mentioned l users in location j with some weight ω_{ij}. For simplicity, we take the weight ω_{ij} to be the number of users in locations i and j who mentioned each other at least once in their tweets. Therefore,

$$\omega_{ij} = \omega_{ji}$$

and the graph generated is undirected, weighted graph.

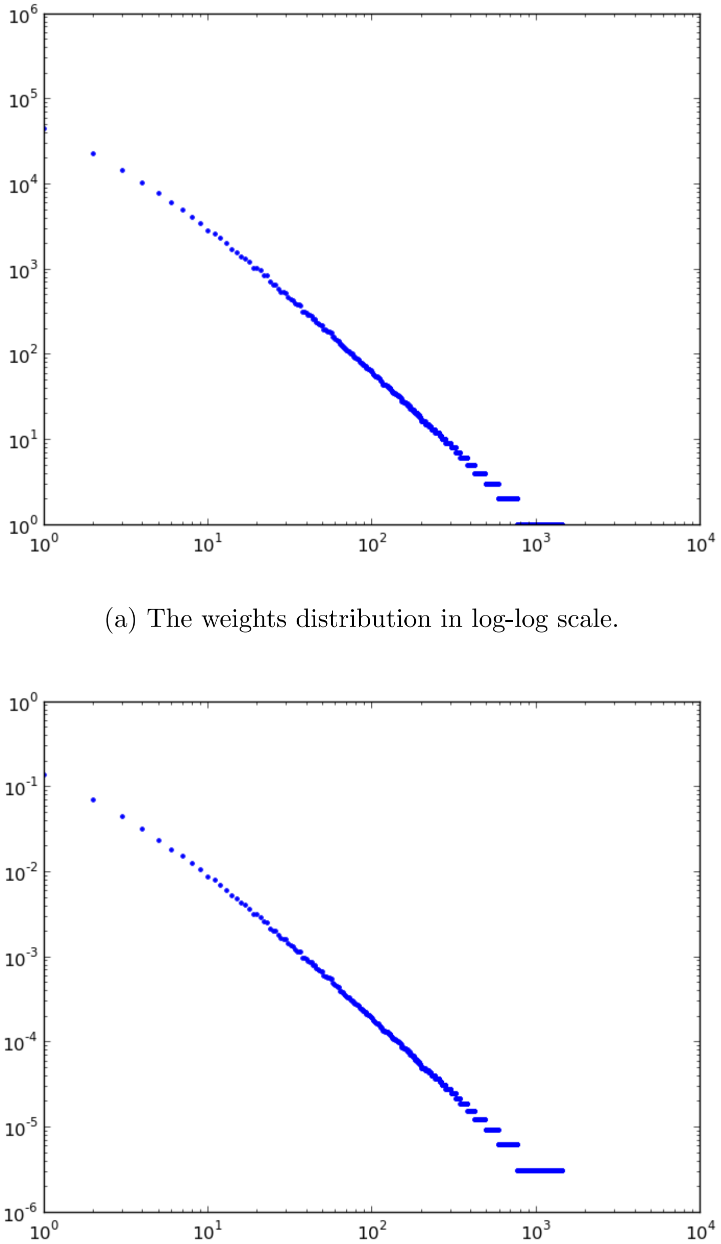
We are interested in constructing graphs based on the strength of the ties between locations which are represented by the graph edges. In this study, we only consider an edge between two locations if the number of common users k exceeds a pre-defined threshold. In this case, we are only considering four different values of the threshold, 5, 10, 15 and 20 which generate four different graphs with varying structure. In the next section we look more closely at the properties of those graphs in order to better understand their overall structure. The properties we are looking at include the basic graph properties such number of nodes and the number of edges and connectivity properties such as the number of connected components and spectral radius. In addition, we also study the distribution of the graph degrees and the eigenvalues of the Laplacian matrix.

Recall that, in linguistics, Zipf's law states: the frequency of any word is inversely proportional to its rank in the frequency table, i.e.

$$f(r) \propto 1/r$$

where r is the rank and $f(r)$ is the frequency. Interestingly, many data distributions in physical and social science follow Zipf's law. Zipf's law can be easily observed if the distribution is plotted in log-log scale where the x-axis is the rank and the y-axis the frequency. In the case where the data follow Zipf's law, the frequencies follow a straight line. Figure 4.3(a) shows the distribution of the weight frequency of the raw data in log-log scale. We see that the plot is almost a straight line which validate Zipf's law. Similarly, Figure 4.3(b) shows the distribution of weight frequency of the normalized data which again follows Zipf's law.

If the data distribution follows Zipf's law, that implies the distribution is fairly predictable. So, in this case, the most frequent weight occurred twice as many times the second most frequent weight had occurred. In general, if the weight w_i is ranked i in terms of frequency, then the value of this frequency f_i equals C/i where C is a universal constant



(a) The weights distribution in log-log scale.

(b) The normalized weight frequency in log-log scale.

Figure 4.3: The distribution of the weights in the global graph is very close to a straight line which validates Zipf's law.

4.4 The graph analysis

In this work, we consider four different graphs based on the number of users who mentioned each other a number of times. The four graphs are constructed based on the number of pairs of users who mentioned each other and this number exceeds a threshold of 5, 10, 15 and 20 respectively. This approach is used to indicate that there is a positive probability that a hashtag mentioned in a location will be mentioned in another location if in the two locations there are users who know each other.

Table 4.1, shows the general structure of the graphs considered. We see that as the threshold increases the number of nodes and the number of edges decrease. Therefore, the graphs are getting less complex as the threshold increases. In order to measure the robustness of the graphs and how it behaves as the value of the threshold increase, we computed the number of connected components, the density and the spectral radius of the four graphs, see Table 4.2. As we can see, as the threshold increases, the number of connected components increases which indicates that the overall robustness of the network decreases. Furthermore, the density and the spectral radius of the graphs decrease as the threshold increase which also implies that the robustness decreases confirming the previous conclusion.

graph	threshold	num. of nodes	max. degree	ave. degree
graph 1	5	6135	1055	26.393
graph 2	10	5140	732	18.6147
graph 3	15	4509	564	15.377
graph 4	20	4149	465	13.2552

Table 4.1: The general structure of the four graphs

Figure 4.4 shows the degree distributions of the four graphs in a log-log plot. Interestingly, all four graphs have very similar degree distributions and all the curves display the same behavior. We notice that the high ranked nodes have relatively similar values of the degrees and the majority of the nodes have low rank.

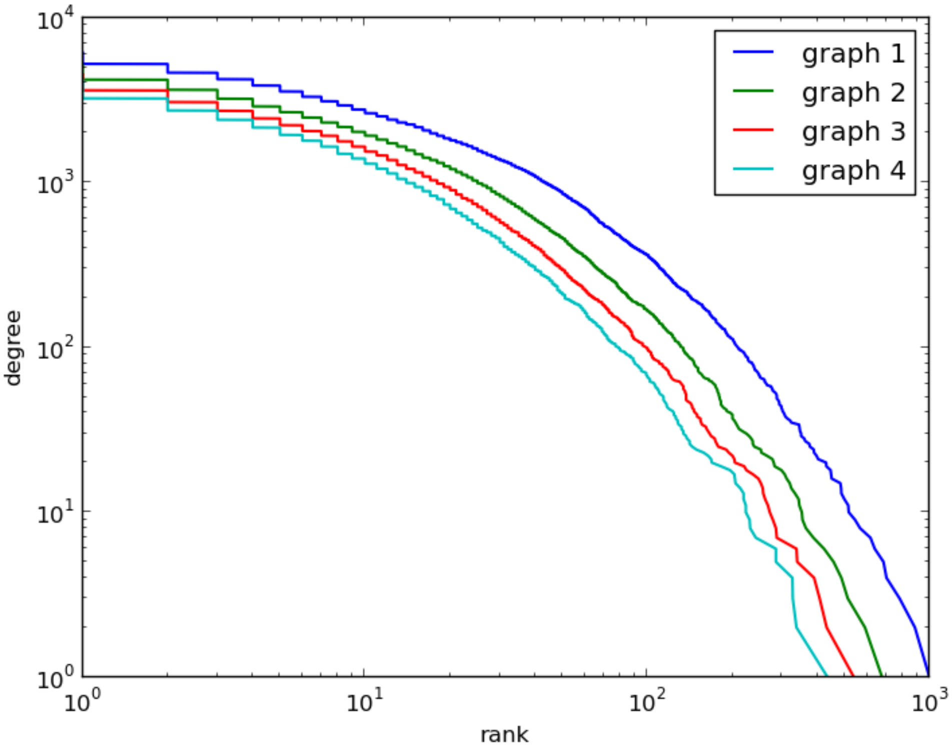


Figure 4.4: The distribution of the degrees of all four graph in log-log plot.

The second experiment is to study the distribution of the eigenvalues of the Laplacian matrix L of the four graphs. Interestingly, for graph 1 and graph 3, the eigenvalues are all non-negative real numbers whereas for graph 2 and 4 have very few complex eigenvalues. For all graphs, we notice that the number of eigenvalues

graph	num. of connected component	density	spec. radius
graph 1	16	0.004303	155.817
graph 2	32	0.00362	106.774
graph 3	36	0.003411	85.873
graph 4	43	0.003196	73.032

Table 4.2: The measures of robustness of the four graphs

close to 1 is much larger than the values of the other eigenvalues, which is clearly an outlier. Notice that, we are only considering the real part of the eigenvalues and it is normalized to be in the interval $[0, 2]$. In addition, the numbers of the smallest and largest eigenvalues are both much larger than the numbers close to them and this is true for all four graphs. Therefore, Figures 4.5, 4.6, 4.7, 4.8 show that all graphs have very similar eigenvalue distribution and the same values for outliers.

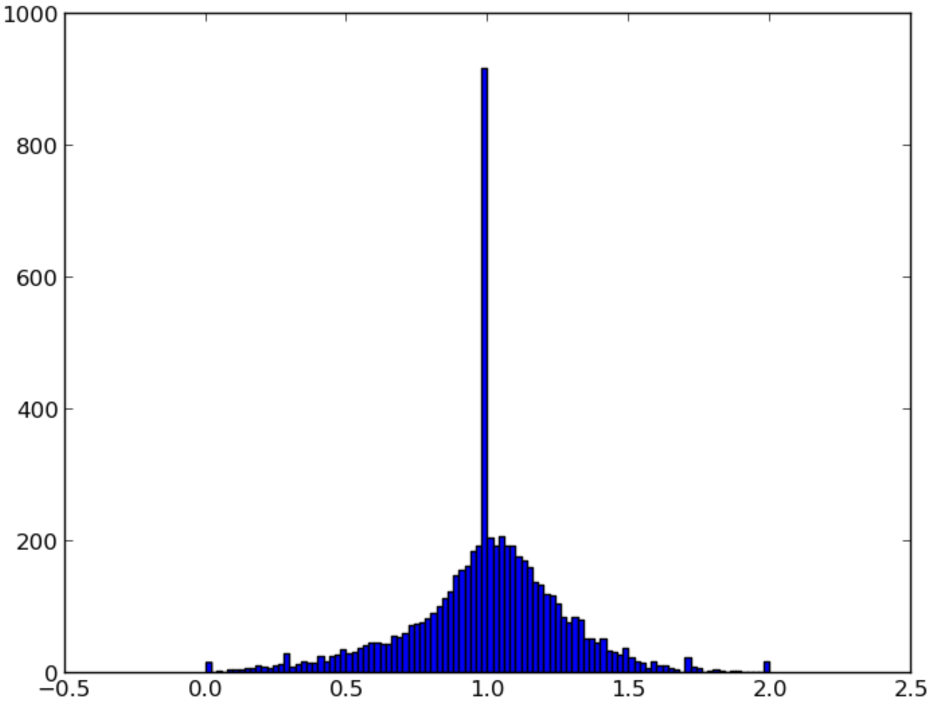


Figure 4.5: A histogram of the eigenvalues of graph 1.

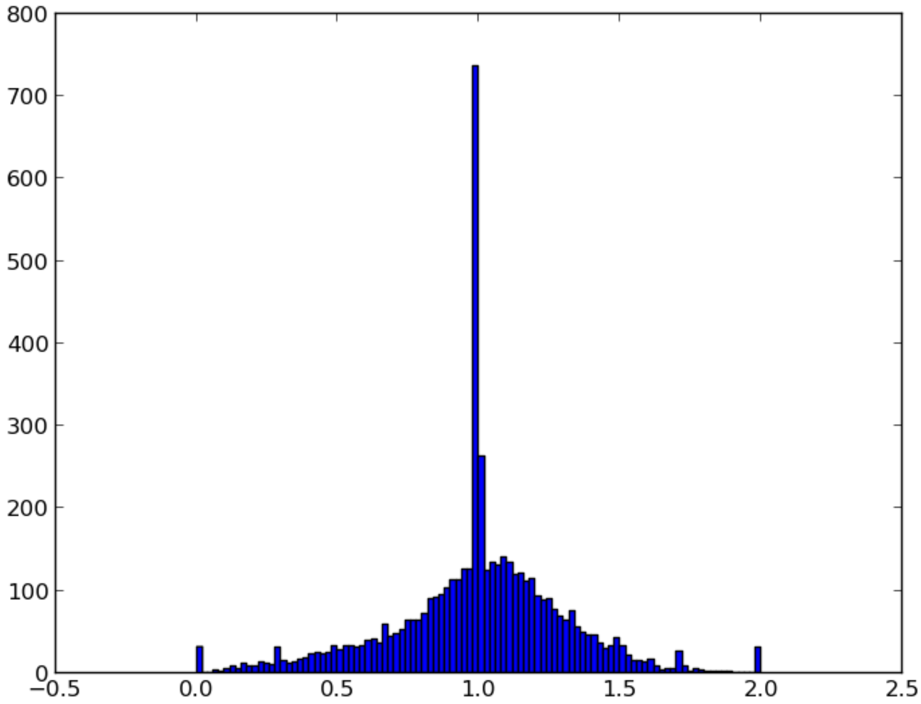


Figure 4.6: A histogram of the eigenvalues of graph 2.

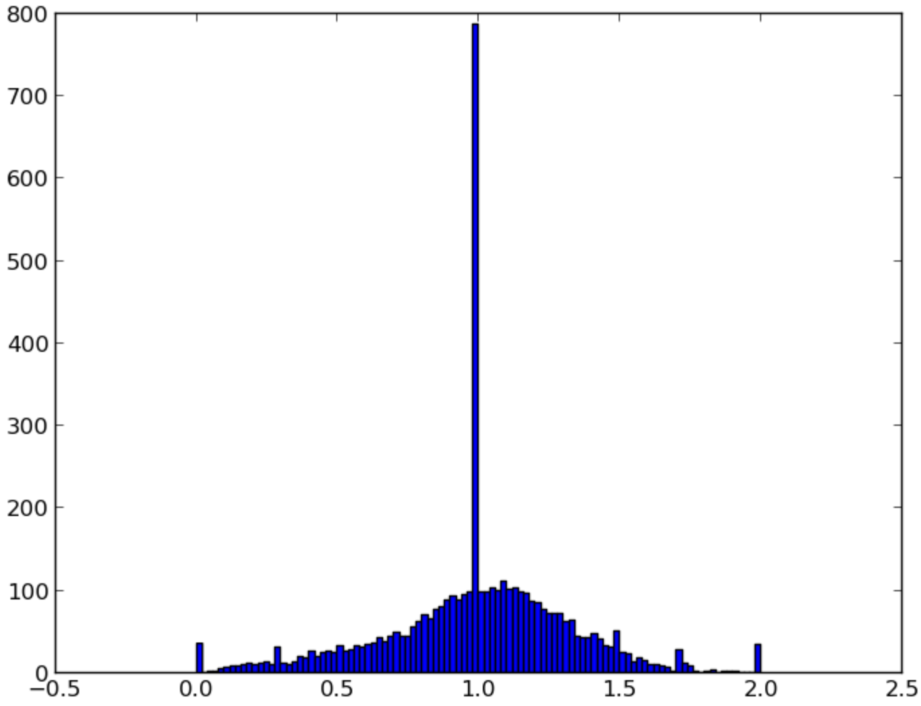


Figure 4.7: A histogram of the eigenvalues of graph 3.

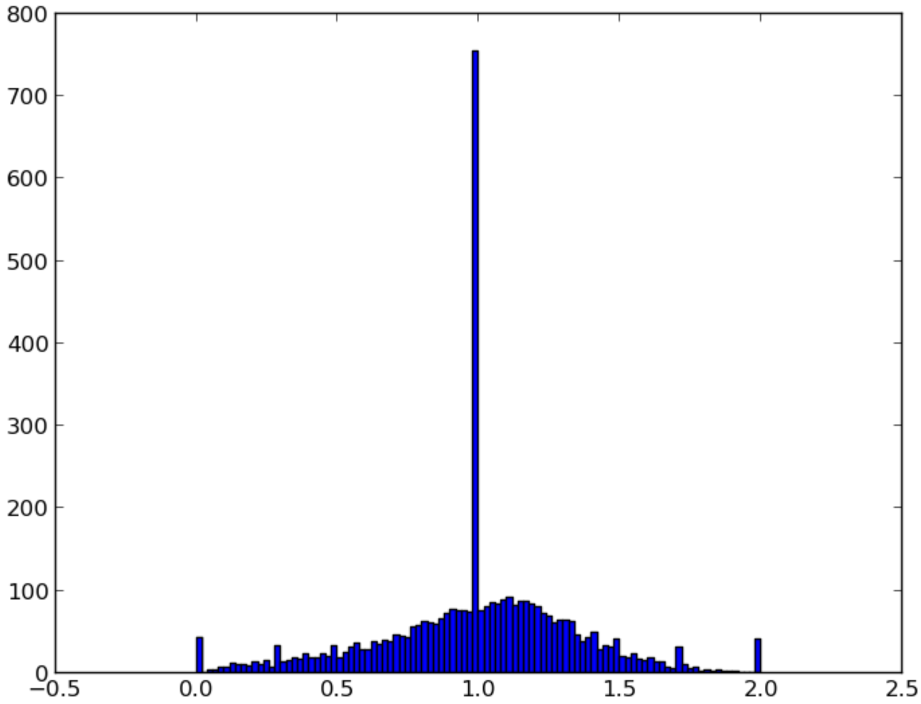


Figure 4.8: A histogram of the eigenvalues of graph 4.

4.5 The influence curve

There is a growing line of research that concerns about the on-line spread of information. In Twitter, information and ideas usually propagate through tweeting, re-tweeting and the use of hashtags. In this work, we study the ways in which the hashtags spread geographically and their propagation patterns. The method we are considering is dividing the world into a large grid of locations and construct a undirected graph based on the “ties” between locations. Hence, there are four graphs have been generated based on the strength of the “ties” between locations. The goal is to investigate the variations in the information diffusion in the different networks.

In order to better understand the geographic propagation of hashtags, we look at the probability of a hashtag to appear in a geographic location after appearing in locations strongly "tied" to that location. Firstly, we need to define what we mean by saying two locations are tied or connected. Secondly, the notion of a location being "exposed" to a hashtag also needs to be defined.

Definition 4.5.1. *We say a location i is k-exposed if at time t there are k neighboring locations where a hashtag H was mentioned before it was mentioned in location i.*

Definition 4.5.2. *We define the influence curve $p(k)$ to be the fraction of locations where the hashtag was adopted directly after their k^{th} exposure to it, given that they had not yet adopted it.*

Given a location i which is a k-exposed to h, the probability that i will adopt H in the future is computed in two ways:

- Ordinal time estimate: here $p(k) = \frac{I(k)}{E(k)}$ where $E(k)$ is the number of locations that were k-exposed to H at some time and $I(k)$ is the number of locations that were k-exposed and used H before becoming $(k+1)$-exposed.

- Snapshot estimate: given interval $T = (t_1, t_2)$, $p(k) = \frac{I(k)}{E(k)}$ where $E(k)$ is the number of locations that were k-exposed to h at time t_1 and $I(k)$ is the number of locations that were k-exposed at time t_1 and used h sometime before t_2.

In this work, we only consider the ordinal time estimate of the influence curve since it requires more detailed data and, therefore, our results are based on this estimate. In order to measure the differences in the hashtags diffusion patterns, we need to introduce the notions of "stickiness" and "persistence". Formally, given the influence function $p : [0, K] \to [0, 1]$, we let

$$R(p) = K \max_{k \in [0,1]} p(k)$$

to be the area of the rectangle with length K and height $\max_{k \in [0,1]} p(k)$. Let $A(p)$ be the area under the curve $p(k)$ assuming k takes non-negative integral values, then

- the persistence parameter $F(p)$ is defined to be

$$F(p) = \frac{A(p)}{R(p)}$$

- the stickiness parameter $M(p)$ is defined to be

$$M(p) = \max_{k \in [0,1]} p(k).$$

The "stickiness" parameter gives us an idea of how large is the probability of adoption given that the location is being "exposed" a number of times. In addition to that, the index of the "stickiness" tells us how many times the location needs to be exposed to the hashtag in order to achieve the maximum probability. The "persistence" parameter, on the other hand, could be viewed as a measure of the influence curve decay. Hence, the smaller the "persistence" value, the smaller the chance the location will adopt the hashtag in the future.

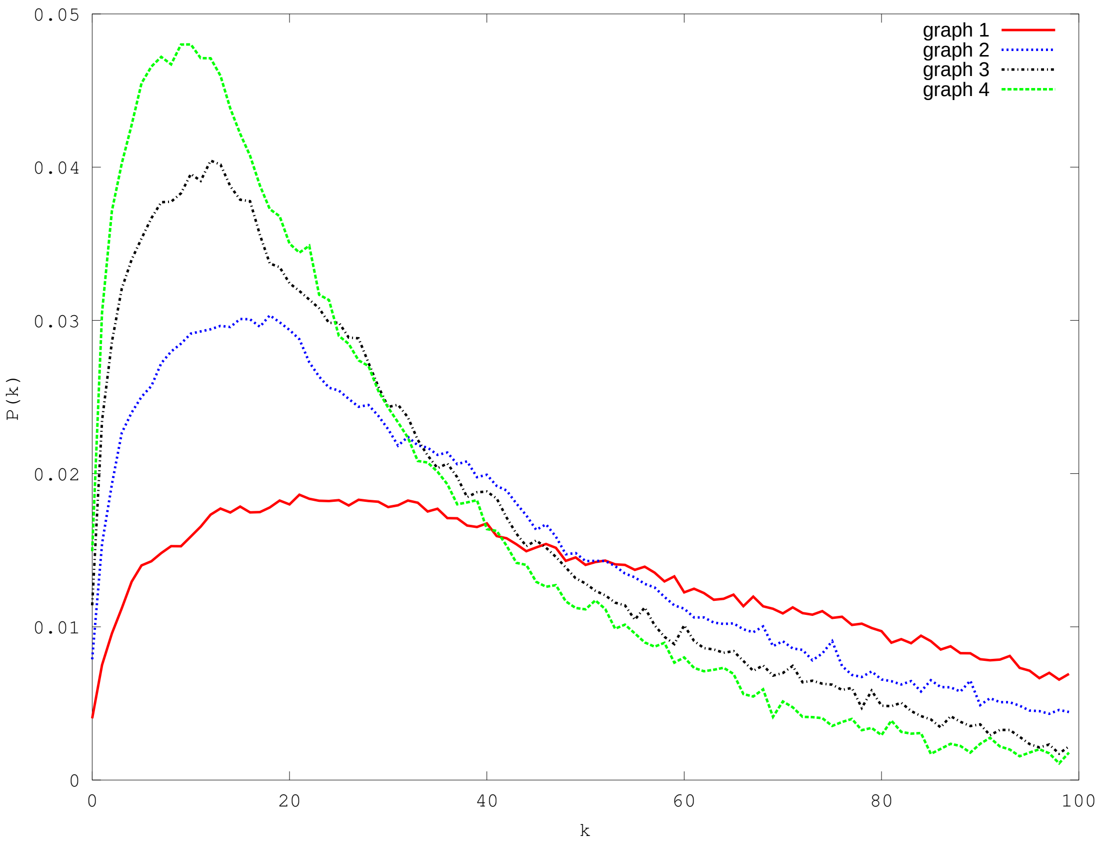


Figure 4.9: A plot of the influence curve of the four graphs. We see that the stronger the "ties", the faster the curve reaches a higher peak. On the other hand, the rate of the decay increases with the strength of connection which validate thes results in Table 4.3.

4.6 Computing influence curve

In this section we describe the methodology we used to compute different influence curves of the different graphs. We consider the top 400 most frequent hashtags and for each one we compute its influence curve. In order to compute the influence curve of a given hashtag, we start by assuming all locations are 0-exposed to the hashtag, then for each location the hashtag appears in, we count the number of neighbors, say k, in which the hashtag has appeared before, then we increment $I(k)$ by 1. The next step is to count the total number of location which were k-exposed at some point in time, i.e. computing $E(k)$ for $k = 1, 2, \dots, K$. Finally we divide $I(k)$ by $E(k)$ for $k = 1, 2, \dots, K$ to find the fraction of locations that adopt the hashtag immediately after being k-exposed. After computing the influence curve for every hashtag, we compute the average of all influence curves.

graph	persistence $F(p)$	stickiness $M(p)$	index k
graph 1	0.71	0.0186	21
graph 2	0.52	0.0303	20
graph 3	0.21	0.0402	17
graph 4	0.18	0.048	16

Table 4.3: The influence parameters of the four graphs.

Figure 4.9 shows the influence curves of all four graphs. We notice that the overall structure of the plots is similar where all curves display similar behavior. We see that all four curves reach the maximum value relatively quickly and then start to decrease at a slower rate to approach zero probability. The difference between the four curves is how fast they reach the value of stickiness and the rate they decrease, see Table 5.3. The “ties” in the graphs are in increasing order and we notice that in Figure 4.9, the stronger the connection between locations, the faster the influence curve reaches the maximum value (or the value of stickiness). On the other hand, a contrasting property can be observed about the persistence: the stronger the “ties” the smaller the value of the persistence. This means that the stronger the “ties” between location, the less the chance of the hashtag to appear in the future.

The second experiment is to divide the hashtags into classes based on their topic. We have chosen four different classes, entertainment, politics, technology and sports.

The four influence curves of the four hashtag categories are shown in Figure 4.10. Since the number of hashtags in each class is much smaller than the total number of hashtags, we notice that the influence curves are not very smooth and there are a lot of sharp oscillations. Therefore, the more hashtags we have, the smoother the influence curve. In this case, we see that the four curves display similar overall behavior. The technology curve reaches the maximum value faster than the rest of curves and the value of the stickiness is higher than the other groups.

We next turn to the challenge of approximating these curves, with some background on Chebyshev polynomials.

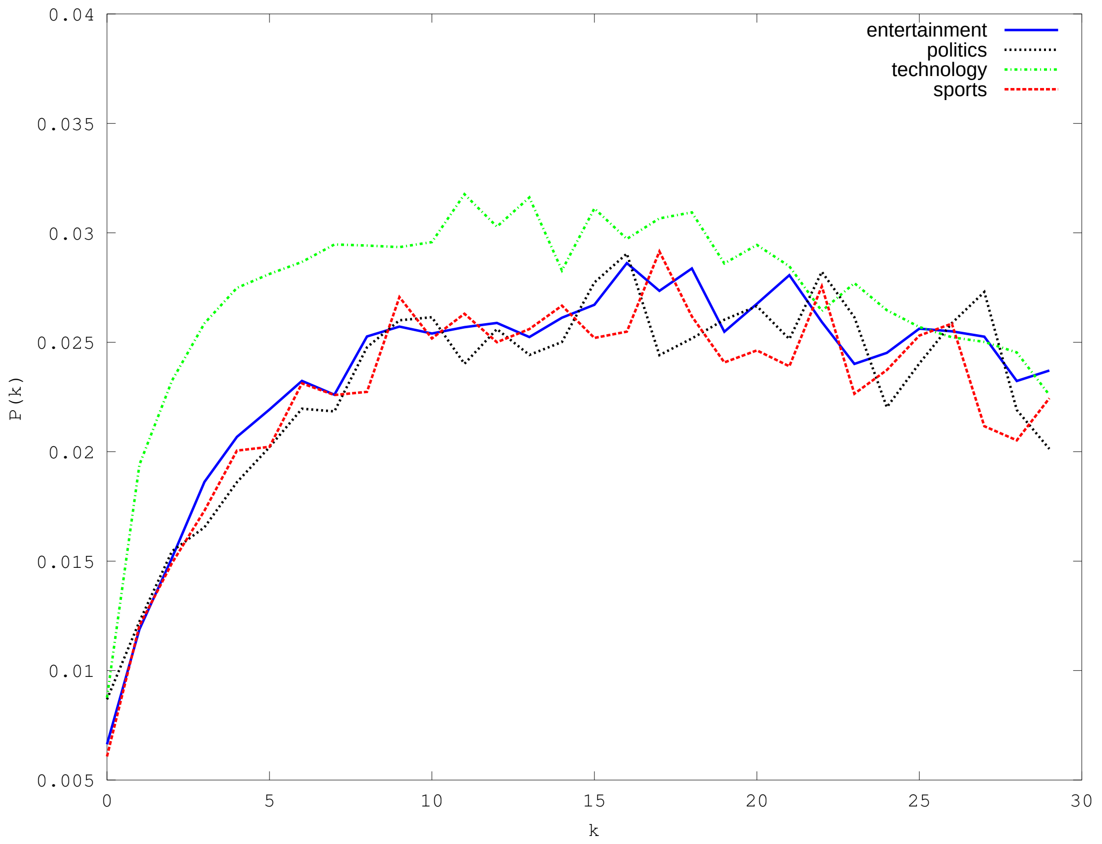


Figure 4.10: A plot of the influence curve of the four classes of hashtags. All curves display similar behavior. The technology curve has a higher value of “stickiness” than the rest of the other classes.

4.7 Chebyshev polynomials

In approximation theory, a set of orthogonal polynomials is a collection of polynomials where the inner product of any two distinct elements is zeros and the inner product of an element with itself is one. Orthogonal polynomials usually form a basis of a subspace of a larger function space . We consider a function $f : \mathbf{R} \to \mathbf{R}$. This function can be approximated by a linear combination of a set of basis functions

$B = \{b_i : \mathbf{R} \to \mathbf{R}, i \in \mathbf{N}\}$ as:

$$f(t) = \sum_{i=0} c_i b_i(t) \quad c_i \in \mathbf{R}$$

In this work, we used Chebyshev polynomials of the first kind as an orthonormal basis B [9, 12]. The n^{th} Chebyshev polynomial $T_n(t)$ is defined as

$$T_n(t) = \cos(n \arccos(t)). \tag{4.1}$$

where $t \in [-1, 1]$.

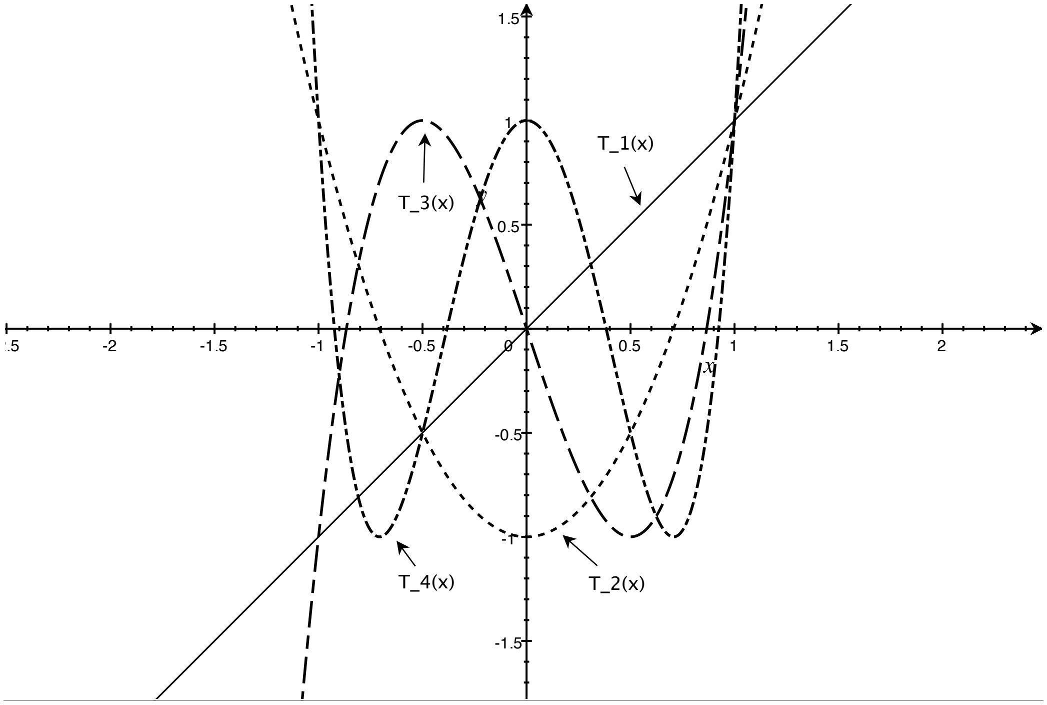


Figure 4.11: A plot of Chebychev polynomials T_1, T_2, T_3 and T_4.

Note that equation 4.1 is in fact defining a polynomial which can be computed using the recursive formula

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x). \end{aligned}$$

Figure 4.11 shows Chebyshev polynomials T_1, T_2, T_3 and T_4 which are orthogonal polynomials of degrees 1, 2, 3 and 4 respectively.

4.8 Chebyshev approximation

Our observation that the more hashtags used in computing the influence curve, the smoother the curve is, since we are taking the average of more and more curves. This leads us to conjecture that the influence curve is converging to a smooth curve which can be computed using simple functions.

Chebyshev polynomials (of the first kind) are a set of orthogonal polynomials and they are widely used in approximation theory. Chebyshev polynomials are the solutions to the Chebyshev differential equations. We use the first six Chebyshev polynomials to approximate the influence curves which are defined as follows:

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ T_5(x) &= 16x^5 - 20x^3 + 5x \end{aligned}$$

By using only six Chebyshev polynomials we managed to approximate the influence curve to a large degree, see Figures 4.12, 4.13. The smooth curve generated by Chebyshev polynomials would approximate the influence curve better if we add more hashtags. In other words, the sequence of influence curves converges to a curve that can be extremely close to Chebyshev curve. Table 4.4 shows the approximate stickiness and the percentage of the error for all four graphs. The error is computed using the mean square error formula and then divided by the total sum of the data (in fact it is the square root of the sum of squares of the data) to get the error percentage.

graph	approx. stickiness	error percentage
graph 1	0.0187	3.5%
graph 2	0.030	5.3%
graph 3	0.038	7.1%
graph 4	0.045	9.5%

Table 4.4: The approximations of the influence curve and the stickiness the four graphs.

From Table 4.4 we notice the error percentage increases as the "ties" between locations increases. On the other hand, the approximation of the stickiness improves as the "ties" between locations decreases, see Table 4.3 for comparison. In other words, we need more data for strongly "tied" graphs to have a better approximation using Chebyshev polynomials. Chebyshev polynomial approximation can be used to obtain a good estimate of the influence curve and its parameters without the need to include a large number of hashtags. In this case we need only six numbers to fully describe the influence curve and that means we do not have to keep all the data to generate the influence curve.

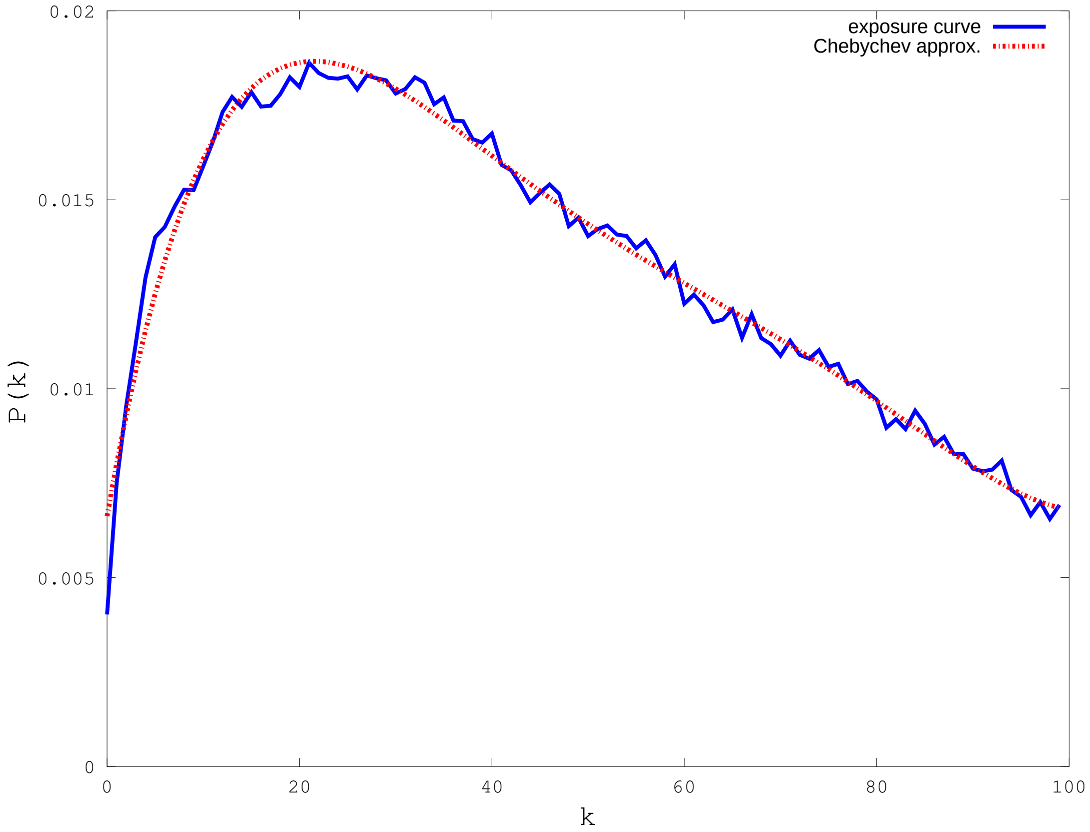


Figure 4.12: The exposure curve of graph 1 and its Chebyshev approximation.

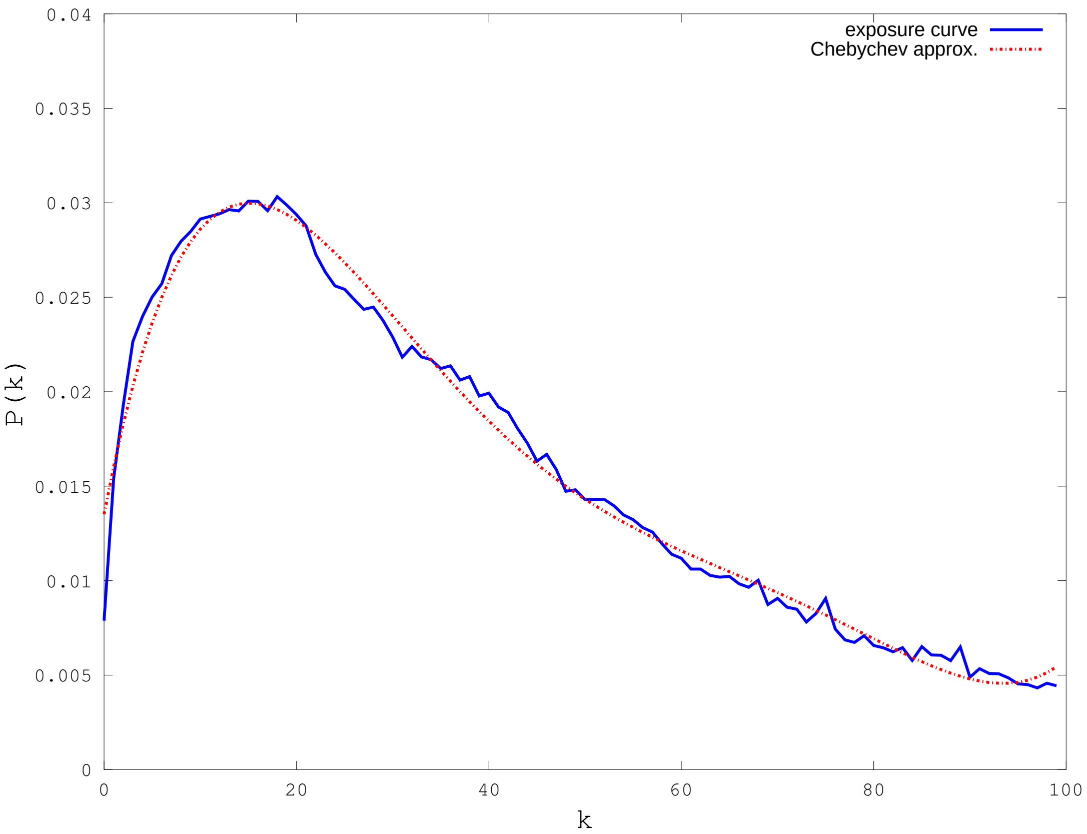


Figure 4.13: The exposure curve of graph 2 and its Chebyshev approximation.

5. CONCLUSION

In this work, we investigated the geographic propagation behavior of hashtags using two different approaches. First we looked at the data as matrix where the rows are the frequencies of the hashtags and the columns are the locations where the hashtag appeared. This form allows us to perform principal component analysis (PCA) in order to reduce the dimensionality of the data space from close to 10,000 to a much smaller number. We discovered considerable redundancy in the data where only 100 variables are needed to retain more than 90% of the information in the dataset.

In the second approach we constructed four different global graphs based the strength of the "ties" between locations. For each graph, we computed the influence curve and its parameters and concluded that the four curves display a similar overall behavior. We noticed that the values of the stickiness of the influence curves increase as the "ties" between locations increases but the persistence of the curve decreases. Finally, we approximated the influence curves using Chebyshev polynomials and we showed that only six polynomials are enough to get a very good approximation of the curve and its parameters.

REFERENCES

[1] Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, pages 58–65. Association for Computational Linguistics, 2011.

[2] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[3] Alan J Izenman. *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer, 2009.

[4] Krishna Y Kamath and James Caverlee. Spatio-temporal meme prediction: learning what hashtags will be popular where. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1341–1350. ACM, 2013.

[5] Krishna Y Kamath, James Caverlee, Zhiyuan Cheng, and Daniel Z Sui. Spatial influence vs. community influence: modeling the global spread of social media. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 962–971. ACM, 2012.

[6] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13,

pages 667–678, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[7] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97, 2010.

[8] DH Maling. Coordinate systems and map projections for gis. *Geographical Information Systems: Principles and Applications. John Wiley & sons*, pages 135–146, 1991.

[9] John C Mason and David C Handscomb. *Chebyshev polynomials*. CRC Press, 2010.

[10] Geerajit Rattanaritnont, Masashi Toyoda, and Masaru Kitsuregawa. Characterizing topic-specific hashtag cascade in twitter based on distributions of user influence. In *Web Technologies and Applications*, pages 735–742. Springer, 2012.

[11] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, 2011. ACM.

[12] Gabor Szegö, Gábor Szegö, and Gábor Szegö. *Orthogonal polynomials*, volume 23. American Mathematical Society New York, 1959.

[13] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth*

ACM international conference on Web search and data mining, pages 643–652. ACM, 2012.

[14] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 599–608. IEEE, 2010.

[15] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.