

Language in Our Time: An Empirical Analysis of Hashtags

Yang Zhang

CISPA Helmholtz Center for Information Security

Saarland Informatics Campus

yang.zhang@cispa.saarland

ABSTRACT

Hashtags in online social networks have gained tremendous popularity during the past five years. The resulting large quantity of data has provided a new lens into modern society. Previously, researchers mainly rely on data collected from Twitter to study either a certain type of hashtags or a certain property of hashtags. In this paper, we perform the first large-scale empirical analysis of hashtags shared on Instagram, the major platform for hashtag-sharing. We study hashtags from three different dimensions including the temporal-spatial dimension, the semantic dimension, and the social dimension. Extensive experiments performed on three large-scale datasets with more than 7 million hashtags in total provide a series of interesting observations. First, we show that the temporal patterns of hashtags can be categorized into four different clusters, and people tend to share fewer hashtags at certain places and more hashtags at others. Second, we observe that a non-negligible proportion of hashtags exhibit large semantic displacement. We demonstrate hashtags that are more uniformly shared among users, as quantified by the proposed hashtag entropy, are less prone to semantic displacement. In the end, we propose a bipartite graph embedding model to summarize users' hashtag profiles, and rely on these profiles to perform friendship prediction. Evaluation results show that our approach achieves an effective prediction with AUC (area under the ROC curve) above 0.8 which demonstrates the strong social signals possessed in hashtags.

CCS CONCEPTS

- Information systems → Data mining; • Human-centered computing → Social networks; Social tagging; Social media; Empirical studies in collaborative and social computing.

KEYWORDS

Hashtag, online social networks, data analysis

ACM Reference Format:

Yang Zhang. 2019. Language in Our Time: An Empirical Analysis of Hashtags. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3308558.3313480>

1 INTRODUCTION

The last decade has witnessed the explosive development of online social networks (OSNs). Leading players in the business, such

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6674-8/19/05.
<https://doi.org/10.1145/3308558.3313480>

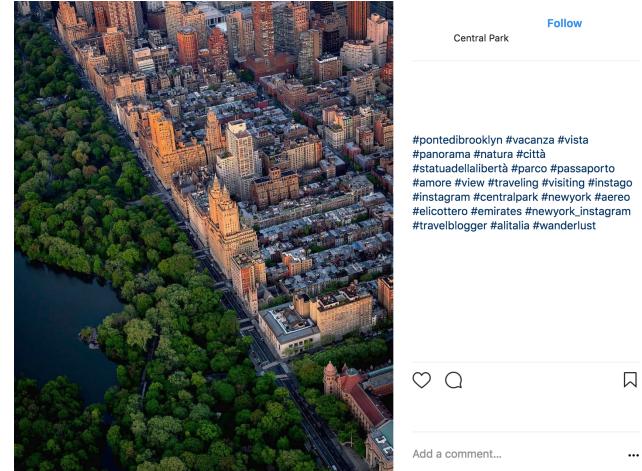


Figure 1: An Instagram post with multiple hashtags.

as Facebook,¹ Twitter,² and Instagram,³ have become the major platform for people to share life moments, communicate with each other, and maintain social relations. Moreover, OSNs have introduced many new notions into our society, such as “like”, “share”, and “check-in”. One particular interesting notion of this kind is *hashtag*.

Created back in 2007, hashtags are designed to help users efficiently retrieve information on Twitter. With development, Instagram has become the major platform for hashtag-sharing. Nowadays, it is very common to see an Instagram post associated with multiple hashtags (see Figure 1 for an example). People also start to use hashtags for various purposes. For instance, many brands use hashtags to promote their products, such as #mycalvins and #shareacoke. Also, hashtags play a major role in various political movements, e.g., #blacklivesmatter and #notmypresident. More interestingly, hashtags have evolved themselves into a new-era language: People have created many hashtags the meanings of which do not exist in the natural language. For instance, #nomakeup attached to a photo indicates that the person in the photo did not wear any makeup; a user publishing #follow4follow means she will follow back others who follow her in OSNs. In another example, #tbt (throwback Thursday) indicates the corresponding photo was taken from old days.

The large quantity of hashtags has provided us with a new way to understand the modern society. Previously, researchers have studied hashtags from various angles [2, 12, 34, 37, 41, 42, 55, 58, 60, 65]. For

¹<https://www.facebook.com/>

²<https://twitter.com/>

³<https://www.instagram.com/>

instance, Souza et al. have analyzed #selfie to understand the online self-portrait convention [58]. Mejova et al. use #foodporn to study people’s dining preference on a global scale [37]. More recently, Zhang et al. investigate the (location) privacy risks stemming from sharing hashtags [65].

Most of these previous works have studied either a certain type of hashtags [34, 37, 41, 42, 58] or a certain property of hashtags [2, 55, 65]. Meanwhile, several general analyses concentrate on hashtags shared on Twitter [12, 60] which is a very different OSN from Instagram with respect to user group, popularity, and functionality [21, 35].

In this paper, we perform the first large-scale empirical study aiming at understanding hashtags shared on Instagram. Our analyses are centered around three research questions summarized from three different dimensions, i.e., the temporal-spatial dimension, the semantic dimension, and the social dimension.

1.1 Research Questions

Different hashtags exhibit different temporal patterns. Holiday-related hashtags, such as #newyear, may have periodic popularity, while the usage of some other hashtags may increase steadily over time. Besides, the information from the spatial dimension may also influence users’ hashtag-sharing behavior: People are more (less) willing to share hashtags when they are at certain places. Therefore, we ask our first research question:

RQ1. *What are the temporal and spatial patterns of hashtags?*

As a new-era language, the semantics of hashtags can uncover many underlying patterns of the modern-style communication. Moreover, due to their inherent dynamic nature, some hashtags may change their meanings within a short time. We therefore ask:

RQ2. *Do hashtags exhibit semantic displacement?*

Following social homophily theory, we hypothesize that friends exhibit more similar hashtag-sharing behavior than strangers. In another way, hashtags possess strong signals for inferring users’ social relations. To test this hypothesis, we ask:

RQ3. *Can hashtags be used to infer social relations?*

1.2 Contribution

We perform the first large-scale empirical analysis of hashtags shared on Instagram. We first sample in total 51,527 Instagram users from three major cities in the English-speaking world, including New York, Los Angeles, and London. We then collect all the posts that these users have shared from the end of 2010 to the end of 2015, and build three separate datasets. In total, our datasets contain more than 41 million Instagram posts shared together with 7 million hashtags.

To address **RQ1**, we first perform clustering to summarize hashtags’ temporal patterns which results in four different clusters. In particular, one cluster of hashtags exhibits strong periodic popularity, some examples in this cluster are #snow, #bbq, and #superbowl. For the spatial dimension, we show that people in all the three datasets tend to share fewer hashtags at certain types of places, such as bars, while more at other types of places, e.g., parks.

For **RQ2**, we first adopt the skip-gram model [38, 39] to map hashtags to continuous vectors and demonstrate that these vectors can very well represent hashtags’ semantics. Relying on the orthogonal Procrustes approach, we measure a hashtag’s semantic displacement between two consecutive years as the distance between its two vectors trained at those years. Evaluation shows that a non-negligible proportion (more than 10%) of hashtags indeed shift their meanings to a large extent. We further define a notion, namely hashtag entropy, to quantify the uniformity of hashtags being shared among users, and observe that hashtags with low entropy are prone to semantic displacement: Correlation coefficients between semantic displacement and entropy in all the datasets are below -0.6.

To answer **RQ3**, we perform a friendship prediction task solely based on hashtags. We propose a bipartite graph embedding approach to learn each user’s hashtag profile and conduct unsupervised friendship prediction based on two users’ profiles’ cosine distance. Extensive experiments show that our approach achieves an effective prediction with AUC (area under the ROC curve) above 0.8 in all the three datasets, and outperforms several baseline models by 20%. This indicates that hashtags indeed possess strong signals on social relations.

To the best of our knowledge, no previous works have studied hashtags’ spatial patterns, semantic displacement, and social signals. We are the first to analyze hashtags from these angles.

We believe our analysis can benefit several parties. The conclusions drawn from answering **RQ1** and **RQ2** can help media campaigns to design more attractive hashtags to engage new customers. The semantic displacement result (from **RQ2**) can help researchers gain a deeper understanding of the OSN culture. The friendship prediction algorithm derived from answering **RQ3** shows that hashtags can also be used as a strong signal for friendship recommendation, which is essential for OSN operators.

1.3 Organization

The rest of the paper is organized as the following. We describe our dataset collection methodology with some initial analyses in Section 2. In Section 3, we investigate the temporal and spatial patterns of hashtags. Section 4 studies the semantic displacement of hashtags. In Section 5, we concentrate on using hashtags to infer social relations. Section 6 discusses the related work in the field. We conclude the paper in Section 7.

2 DATASETS AND INITIAL ANALYSES

In this section, we first describe our data collection methodology, then perform some initial analyses on hashtags.

2.1 Datasets Collection

We resort to Instagram to collect our datasets for experiments. Launched in October 2010, Instagram is a social network service concentrating on photo sharing. By now, it is the second most popular OSN with more than 1 billion monthly active users.⁴ Instagram is the major social network for hashtag-sharing (Figure 1),

⁴<https://instagrampress.com/our-story/>

Table 1: Summary statistics of all the three datasets.

	New York	Los Angeles	London
No. posts	20,673,946	11,907,967	8,640,637
No. hashtags	4,095,575	3,071,158	1,702,675
No. check-ins	1,609,062	883,862	586,420
No. users	25,735	14,687	11,105
No. social links	82,964	36,434	12,900

many popular hashtags are strongly related to Instagram itself, e.g., #instagood and #instamood.

We collect our dataset relying on Instagram’s public API on April, 2016. The first step is finding a sample of users [40]. In the literature, there exist multiple methods for this task. One is relying on the OSN’s streaming API [46] which Instagram does not provide. Another way is generating random integers and query the Instagram API to see whether these numbers are valid user IDs [58]. However, this approach needs us to further investigate how Instagram users’ IDs are distributed. In this paper, we instead sample users based on their locations following previous works [61, 65].

We concentrate on three major English-speaking cities including New York, Los Angeles, and London. In the first step, we query the API of Foursquare,⁵ a location-based social network, to find all the Foursquare location IDs in the three cities. Then, we use Instagram’s API to map all the Foursquare location IDs to the corresponding Instagram location IDs.⁶ Next, we query all the obtained Instagram location IDs to get all users who have ever shared posts at those locations. As in previous works, we further perform some preprocessing to filter out the users matching any of the following criteria [3, 7, 44, 45, 57].

- users with less than 20 check-ins in each city
- users whose numbers of followers are above the 90th percentile (celebrities) or below the 10th percentile (bots)
- users not using human images in their profile photos⁷

In total, we obtain 51,527 Instagram user IDs. Then, we collect all these users’ Instagram posts from the creation time of their accounts until December 31st, 2015. Each post is organized in the following format.

$$\langle \text{user ID}, \text{time}, \text{hashtags}, \text{location ID} \rangle$$

Note that a post is not necessarily associated with a set of hashtags or a certain location ID. We further query Instagram’s API to extract users’ social relationships. We consider two users to be friends if they follow each other [3, 7, 11].

We treat datasets collected from users in the three cities separately to ensure the robustness of our analyses. In total, the New York dataset (dataset collected from users sampled in New York) contains more than 20 million Instagram posts, the Los Angeles dataset contains 11 million posts, and the London dataset contains 8.6 million posts. Moreover, the three datasets contain more than 7

⁵<https://foursquare.com/>

⁶Instagram’s API was connected with Foursquare’s API until April 20th, 2016 (<https://www.instagram.com/developer/changelog/>).

⁷This is done with the help of Face++’s API.

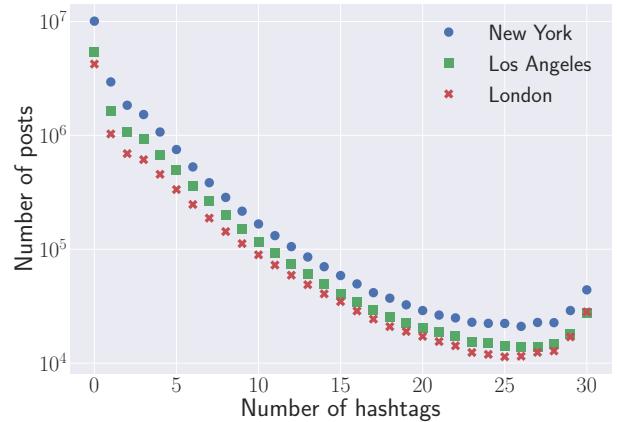


Figure 2: Distribution of the number of hashtags in each post in three datasets. The y-axis is in log scale. The proportions of posts with no hashtags are 48.69% in the New York dataset, 45.28% in the Los Angeles dataset, and 48.92% in the London dataset.

Table 2: Top 10 hashtags with the highest share times in three datasets.

New York	Los Angeles	London
#nyc	#love	#london
#love	#losangeles	#love
#tbt	#tbt	#instagood
#nofilter	#california	#travel
#brooklyn	#instagood	#summer
#latergram	#nofilter	#photooftheday
#instagood	#foodporn	#food
#art	#family	#art
#travel	#art	#instadaily
#summer	#fun	#architecture

million hashtags. Table 1 presents some summary statistics of the datasets.

Ethical Considerations. Our data collection is done through Instagram’s public API in 2016. All the datasets are stored in a central server with restricted access. We further anonymize the datasets by removing all users’ screen names, and replacing their Instagram IDs with randomly generated numbers. Our experiments are conducted on these anonymized datasets.

2.2 Initial Analyses

Figure 2 depicts the distribution of the number of hashtags in each post. We observe that more than half of the posts are associated with at least one hashtag in all the datasets. Moreover, there are around 30% posts associated with 1 to 3 hashtags. The small increase close to 30 is due to the fact that Instagram imposes an upper bound of 30 hashtags per post.⁸

⁸<https://help.instagram.com/351460621611097>

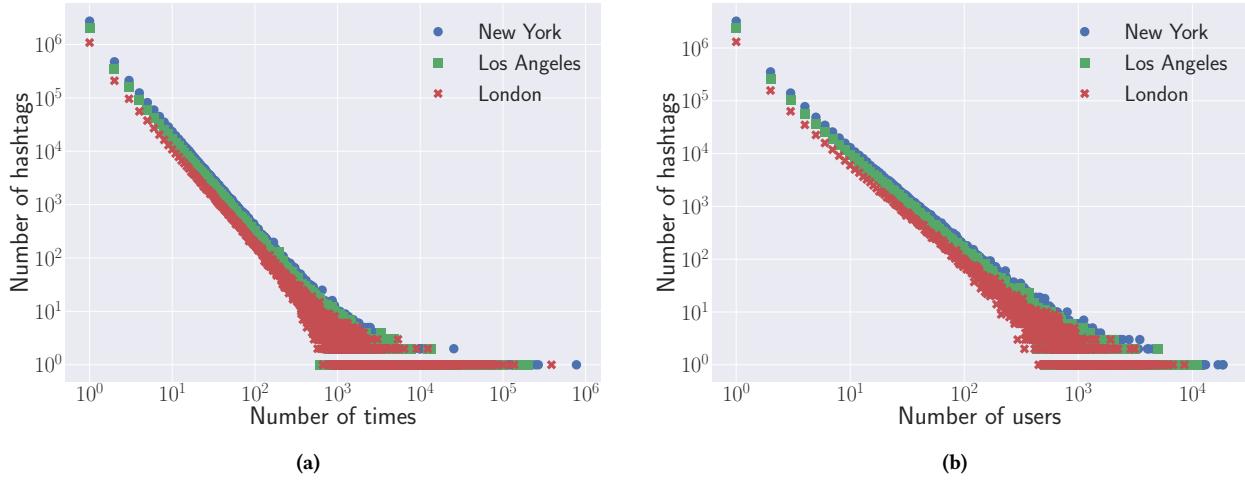


Figure 3: (a) Distribution of the number of times a hashtag is shared in three datasets; (b) Distribution of the number of users each hashtag is used by in three datasets. Both x and y axes are in log scale.

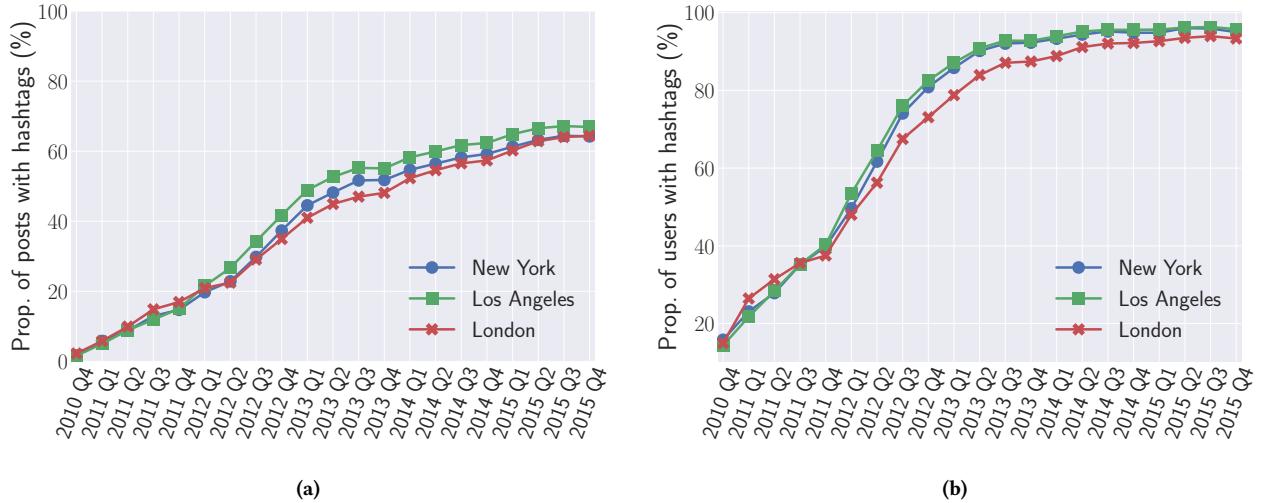


Figure 4: (a) Proportion of posts that are associated with hashtags from 2010 Q4 to 2015 Q4 in three datasets; (b) Proportion of users that use hashtags from 2010 Q4 to 2015 Q4 in three datasets. Q1 represents the first quarter of a year.

Figure 3a plots the distribution of the number of times each hashtag is shared, referred to as share times, while Figure 3b plots the distribution of the number of users each hashtag is used by. As expected, both distributions follow power law, i.e., most of the hashtags are shared only a few times and by a small number of users. Table 2 lists the hashtags with the highest share times. As our users are sampled by cities, many of these popular hashtags are related to city names, such as #nyc, #losangeles, and #london. However, general popular hashtags are captured as well, e.g., #love, #tbt, and #instagood. This indicates that our datasets are suitable for conducting the study.

Figure 4 depicts hashtags' general temporal patterns from 2010 until 2015. We see that in the fourth quarter of 2010 (2010 Q4) when

Instagram was launched, there are less than 2% of the posts associated with hashtags, after 5 years (2015 Q4), the number becomes almost 70% (Figure 4a). Within the same time period, the proportion of users that ever use hashtags grow from 18% to 97% (Figure 4b). These results fully demonstrate the popularity of hashtags.

3 HASHTAGS IN THE TEMPORAL-SPATIAL DIMENSION

This section concentrates on our first research question: *What are the temporal and spatial patterns of hashtags?* We start by investigating the temporal patterns of hashtags, then discuss the relation between hashtags and locations.

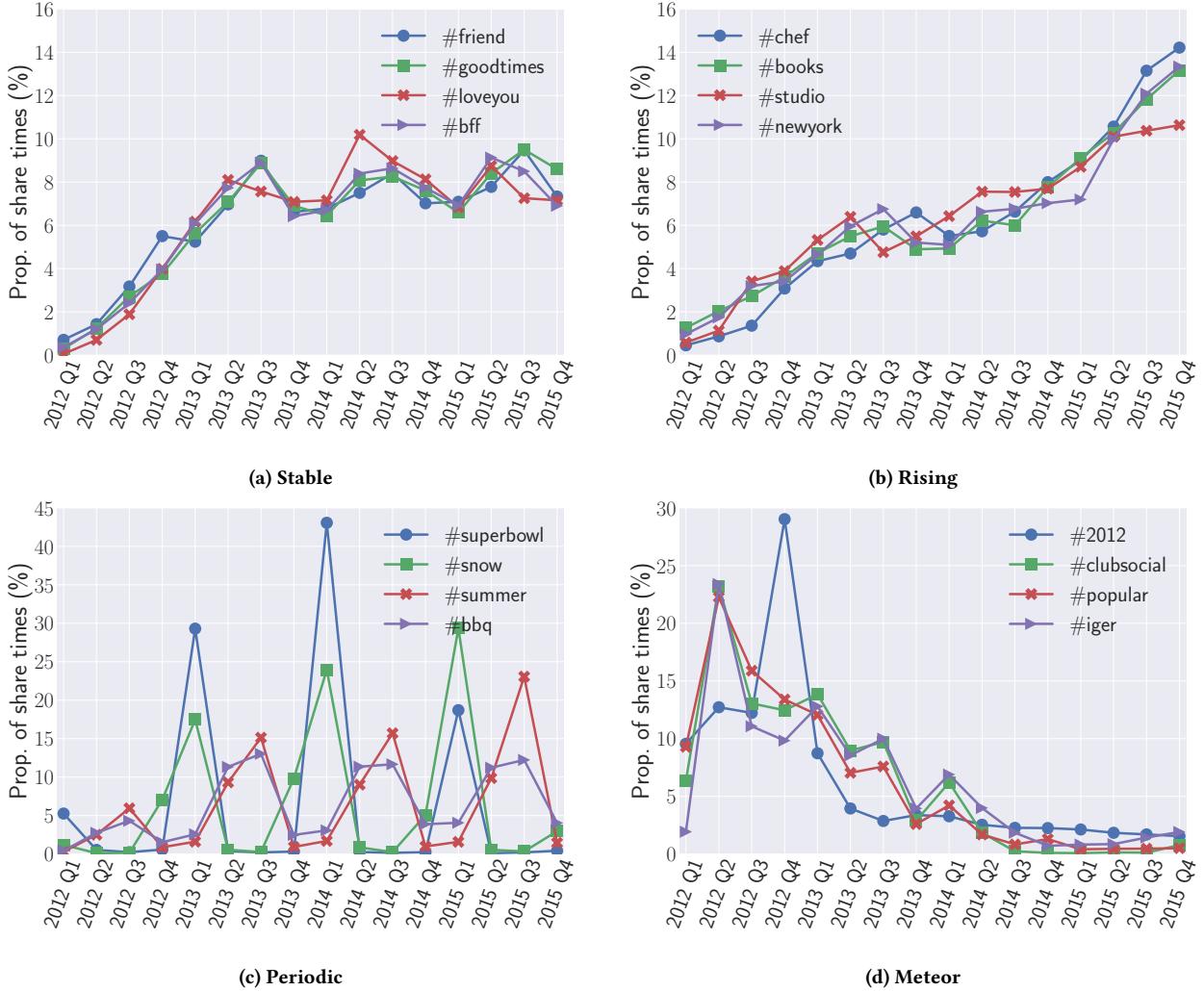


Figure 5: Temporal patterns of some example hashtags belonging to different temporal clusters in the New York dataset.

3.1 Temporal Patterns

We hypothesize that different hashtags exhibit different temporal patterns. Some hashtags should have periodic patterns, such as holiday-related ones, while others' share times may increase over time. Besides, there may exhibit other temporal patterns for hashtags. To perform a quantitative study, we resort to machine learning clustering to discover different temporal patterns.

We aggregate each hashtag's proportion of share times to the granularity of quarters starting from the first quarter in 2012 to the last quarter of 2015.⁹ This indicates that each hashtag's temporal pattern is organized as a 16-dimension vector (4 years \times 4 quarters). k -means is adopted to perform clustering. We first try to fit each hashtag's temporal patterns, i.e., the 16-dimension vector, directly to k -means, however, the resulting clusters are not very promising.

⁹We neglect the data in 2010 Q4 and 2011 due to the small quantity as shown in Figure 4.

This is due to the fact that some hashtags may share similar temporal patterns, however, their patterns' peak happen at different time, e.g., #christmas and #halloween.

Instead, we manually define features over each hashtag's temporal patterns which are described in Table 3. To ensure the robustness of our results, we concentrate on the top 1,000 hashtags in terms of share times [19]. In order to select a suitable number of clusters for k -means, i.e., k , we adopt the Silhouette value. Experimental evaluation suggests that $k = 4$ leads to the highest Silhouette value (around 0.6) in all the three datasets.

The resulting four clusters are named *Stable*, *Rising*, *Meteor*, and *Periodic* based on their temporal patterns. Their detailed descriptions are as follows.

- **Stable:** Hashtags in the first cluster in the beginning exhibit an increase in their usage. After a certain time point, their share times become stable.

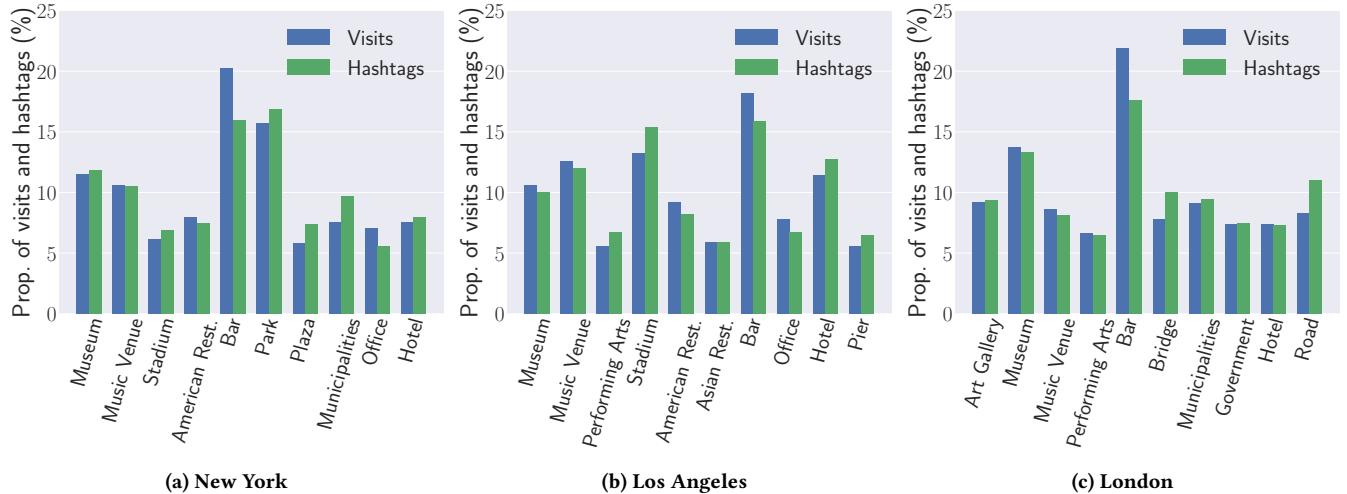


Figure 6: Proportions of visits and hashtags shared at the most popular location categories in three datasets. Rest. in (a) and (b) means Restaurant.

Table 3: Features defined over each hashtag h 's temporal pattern \mathcal{T}_h .

Description
Standard deviation of \mathcal{T}_h
Largest 3 values in \mathcal{T}_h
Mean of the 3 largest values in \mathcal{T}_h
Standard deviation of the 3 largest values in \mathcal{T}_h
Standard deviation of the 3 largest values' indices in \mathcal{T}_h
Smallest 3 values in \mathcal{T}_h
Mean of the 3 smallest values in \mathcal{T}_h
Standard deviation of the 3 smallest values in \mathcal{T}_h
Standard deviation of the 3 smallest values' indices in \mathcal{T}_h

- Rising: The share times of hashtags in this cluster increase steadily over time.
- Meteor: In this cluster, hashtags at some temporal point are suddenly shared a large number of times.
- Periodic: Hashtags in this cluster exhibit periodic popularity.

The sizes of the four clusters are not uniform. The Rising cluster is the largest one containing around 60% of all hashtags, followed by the Stable cluster (around 25%). Meteor, on the other hand, is the smallest cluster with around 7% of all hashtags.

Figure 5 depicts the temporal patterns of some hashtags in different clusters in the New York dataset. We can make several interesting observations. For instance, the Periodic cluster contains not only hashtags about seasons or holidays, such as #summer, but also those about season-related activities, such as #bbq. Moreover, hashtags describing (annual) sports events belong to this cluster as well, e.g., #superbowl. As expected, hashtags that are specific to a certain point of time are categorized in the Meteor cluster, one example is #2012. Interestingly, some hashtags that are specific to Instagram also belong to the Meteor cluster. For instance, #iger in

the second quarter of 2012 gains a large popularity, while in 2015, no one uses it anymore.

3.2 Spatial Patterns

Next, we investigate the spatial patterns of hashtags. In particular, we are interested in at which types of places people are more willing to share hashtags. To get the type/category of a location in an Instagram post, we again resort to Foursquare (see Section 2). Foursquare organizes all its location categories into a two-level tree,¹⁰ we focus on the fine-grained second-level which contains more than 300 different location types.

We pick the top 10 location categories that users visit the most number of times in each dataset within the corresponding city, and calculate the proportion of hashtags shared among these categories. For comparison, we further calculate the proportion of users' visits to locations under these categories.

Figure 6 plots the results. First of all, places belonging to the Bar category are among the most popular locations for people to visit in all the datasets. However, the proportions of hashtags shared at bars drop in all the cases, e.g., the drop in the London dataset is around 5%. This means people are less likely to share hashtags at bars. Another type of locations exhibiting the same result is Office. On the other hand, people are more willing to share hashtags at outdoor places, e.g., Park in New York, Pier in Los Angeles, and Road in London. These results show that the spatial information indeed influences users' hashtag-sharing behavior.

4 SEMANTIC CHANGE

In this section, we address our second research question: *Do hashtags exhibit semantic displacement?* We start by describing how to express each hashtag's semantics, then focus on semantic displacement.

¹⁰<https://developer.foursquare.com/docs/resources/categories>

Table 4: 10 hashtags that are most semantically similar to #family, #sushi, and #r2d2 in the New York dataset.

#family	#sushi	#r2d2
#familytime	#sashimi	#c3po
#cousins	#nigiri	#droids
#aunt	#sushiporn	#bb8
#father	#spicytuna	#artoo
#mom	#chirashi	#astromech
#sisters	#sushiroll	#carbonite
#grandparents	#unagi	#xwing
#nephews	#sushirolls	#obiwankenobi
#familylife	#shrimptempura	#lukeskywalker
#siblings	#californiaroll	#starwars

4.1 Semantics of Hashtags

As a new-era language, hashtags convey interesting meanings. To study this, our first step is finding a tool to represent each hashtag’s semantics. Here, we adopt the skip-gram model with negative sampling [38, 39]. Skip-gram, designed following the distributional hypothesis in linguistics, is essentially a shallow neural network model, it maps each word into a continuous vector which preserves the information of the word’s contexts words in a large corpus.

In our case, we treat each hashtag as a “word”, and all hashtags in a post as one “sentence”. Then, we perform skip-gram over all the “sentences” for each dataset. We set each hashtag’s vector’s dimension to 300 following previous works [19, 38, 39].

To show the effectiveness of skip-gram on capturing hashtags’ semantics, we perform a qualitative study to find some hashtags’ most semantically similar ones with respect to the shortest cosine distance between learned vectors. Table 4 presents the results for #family, #sushi, and #r2d2 in the New York dataset. As we can see, all the semantically similar hashtags found for these hashtags indeed express quite similar meanings, e.g., #r2d2 is similar to #c3po, #droids, and #bb8.

4.2 Semantic Displacement

With the cultural evolution, some words in natural language exhibit semantic displacement, e.g., the word “gay”. We are interested in whether semantic displacement can be observed on hashtags as well. Moreover, due to hashtags’ inherent dynamic nature, their semantic displacement should be much faster than that of words which normally takes decades, therefore, our datasets spanning over 5 years are sufficient for this study.

To measure semantic displacement of each hashtag, we first split our datasets by years starting from 2011 to 2015. Then, we perform skip-gram to map hashtags in each year into a vector. This means each hashtag has up to 5 vectors. A *single semantic displacement* of a hashtag is measured over two consecutive years as the cosine distance of the hashtag’s two vectors at those years. The *overall semantic displacement* of a hashtag is the mean of all its single semantic displacements.

However, the stochastic nature of the skip-gram model drives vectors learned at different years to reside at different coordinate axes. Therefore, we need to align each hashtag’s vectors at different

Table 5: 10 hashtags with the highest overall semantic displacement in three datasets.

New York	Los Angeles	London
#ontheroad	#hashtag	#iphoneography
#epic	#epic	#insta
#ilovemyjob	#hipstamatic	#world
#free	#winning	#today
#starwars	#nofilter	#popular
#winning	#insta	#potd
#pictureoftheday	#ig	#walking
#nofilter	#iphone	#ignation
#iphonography	#free	#instamoment
#yolo	#random	#studio

Table 6: 10 hashtags with the lowest overall semantic displacement in three datasets.

New York	Los Angeles	London
#nyc	#sweettooth	#pet
#beach	#dessert	#trees
#sky	#sweets	#beach
#sand	#trees	#tube
#eastriver	#palmtrees	#cat
#leaves	#soup	#kitty
#waves	#ocean	#skyscraper
#salad	#sunset	#thames
#sunset	#building	#smoothie
#pasta	#foodporn	#glutenfree

years together before calculating their cosine distance [19, 25]. To this end, we apply the orthogonal Procrustes approach used by Hamilton et al. [19]. Concretely, we use $M^{(t)} \in \mathbb{R}^{300 \times |\mathcal{H}|}$ to represent the matrix containing all the hashtags’ vectors’ at year t , and \mathcal{H} to represent the set of all hashtags. Then, we solve the following optimization objective function

$$\arg \min_X ||XM^{(t)} - M^{(t+1)}||_F$$

subject to $XX^T = I$

by applying singular value decomposition (SVD). Here, $|| \cdot ||_F$ denotes the Frobenius norm. After obtaining the result, we align vectors of a hashtag in two consecutive years together.

Again, to ensure the robustness of our results, we concentrate on the top 1,000 hashtags that are shared the most number of times. Our experimental evaluation shows that some hashtags indeed exhibit a large semantic displacement. In our three datasets, there are at least 10% of the hashtags with more than 0.4 overall semantic displacement. Given the fact that cosine distance lies in the range of $[0, 2]$,¹¹ the semantic shift is pretty large. Table 5 and Table 6 list the hashtags with the highest and lowest overall semantic displacement. We see that hashtags that are specific to Instagram have quite large semantic displacement, such as #insta, #nofilter, and #ig. Many of

¹¹Vectors learned by skip-gram are in \mathbb{R}^{300} , thus, the cosine distance is in $[0, 2]$.

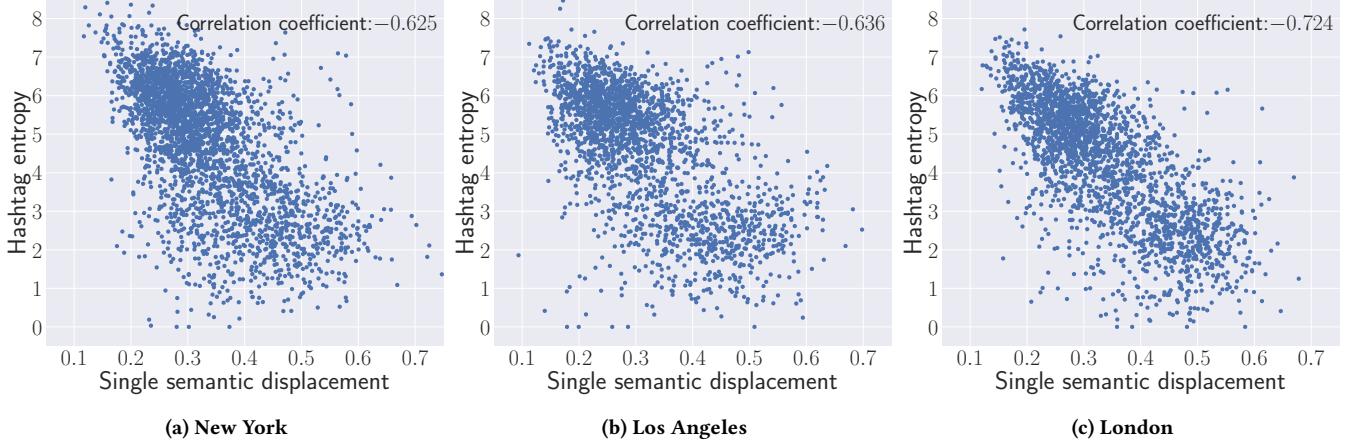


Figure 7: Relation between single semantic displacement and hashtag entropy in three datasets. Hashtag entropy is calculated on a yearly-base.

these hashtags are initially created to increase Instagram posts' popularity. We conjecture that when people are more used to these hashtags, they cannot attract as much attention as before, thus start to convey different meanings. On the other hand, we observe from Table 6 that nature-related hashtags are among those whose semantics stay rather stable, e.g., #sky, #trees, and #ocean. Also, food-related hashtags, such as #salad, #soup, and #smoothie, do not change their meanings that much.

Hamilton et al. show that high frequency words are less likely to change their meanings [19]. However, this is not the case for hashtags. We perform a correlation analysis between hashtags' frequency, i.e., share times, and single semantic displacement, and obtain a rather weak correlation (correlation coefficients are around -0.2 in all the datasets). Instead, we hypothesize that hashtags that are more uniformly shared among users are less likely to shift their meanings.

We propose *hashtag entropy* to measure each hashtag's sharing uniformity among users. Formally, a hashtag h , its entropy is defined as:

$$-\sum_{u \in \mathcal{U}} p(u, h) \log p(u, h)$$

where \mathcal{U} denotes the set containing all the users and $p(u, h)$ represents the proportion of h shared by user u . It is worth noting that our hashtag entropy follows the same definition as the Shannon entropy: Higher entropy implies that the hashtag is more uniformly shared among users.

For each hashtag, we calculate multiple entropy at each year (from 2012 to 2015), then correlate each entropy with the corresponding single semantic displacement. Figure 7 depicts the results. In all the three datasets, we obtain strong negative correlation. In particular, the correlation coefficient in the London dataset is -0.724. These results show that if a hashtag is shared by many users in a similar frequency, i.e., high entropy, then it is less prone to semantic displacement.

5 HASHTAGS AND FRIENDSHIP

In this section, we address our last research question: *Can hashtags be used to infer social relations?*. We first describe our methodology on using hashtags for friendship prediction, then present the experimental evaluation.

5.1 Methodology

Social network is the major platform for people to share hashtags, this naturally leads to the question whether hashtags are related to people's social relations. To answer this question, we perform a friendship prediction task solely based on hashtags.

Friendship prediction is normally modeled as a binary classification task [1, 3, 4, 29, 31, 48, 57]. We can manually define features over two users' hashtags, and perform prediction either in a supervised manner (with the help of machine learning classifier) or an unsupervised manner. However, feature engineering is very time-consuming, and in many cases, the resulting features are not complete. Also, to define features over two users' hashtags, a natural approach is concentrating on their common hashtags. As most pairs of users share no common hashtags, this approach can be only applied to a small subset of user pairs, which cannot provide a complete picture on the relation between hashtags and social networks.

Instead, we choose to learn a hashtag profile, i.e., a feature vector, for each user, and compare two users' profiles to perform friendship prediction. As all users' hashtag profiles (feature vectors) are in the same dimension, this allows us to predict any pair of users' friendship.

Our hashtag profile learning follows the recent advancement of graph embedding [3, 13, 17, 20, 22, 32, 47, 50, 51, 53, 59, 62, 64]. In this setting, profiles are automatically learned following a general optimization objective function. Concretely, we organize users and hashtags into a weighted bipartite graph. For an edge connecting a user and a hashtag, its weight equals to the number of times the user shares the hashtag. We simulate a certain number of random walks starting from each user, referred to as *walk times*, on the

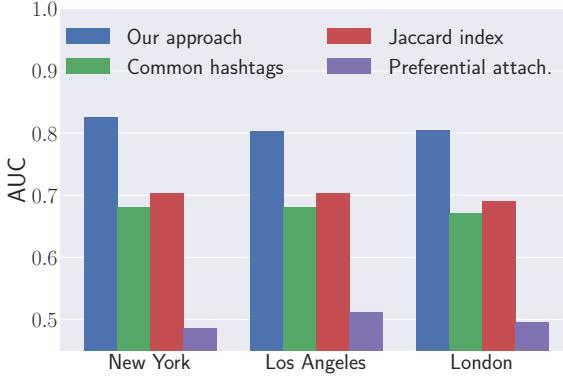


Figure 8: AUC for friendship prediction in three datasets. Preferential attach. means preferential attachment.

graph. The transition probability from each node to the next one follows the corresponding edges’ weight. Each walk has a certain length, specified by the number of steps, which we refer to as *walk length*. This leaves us with a set of random walk traces. Then, we rely on the following optimization objective function to learn each user’s hashtag profile.

$$\arg \max_{\theta} \prod_{v \in \mathcal{U} \cup \mathcal{H}} p(v|N(v); \theta)$$

Here, $N(v)$ denotes the neighborhood of node v ¹² and $\theta(v)$ is the learned profile of node v . Moreover, $p(v|N(v); \theta)$ is modeled with a softmax function. Different from previous graph embedding approaches which define their objective functions following skip-gram [17, 47], our objective function is essentially the continuous bag-of-words (CBOW) model [38, 39]. We choose CBOW over skip-gram due to its better performance. In addition, we apply the negative sampling approach to speed up the learning process.

In the end, for any two users, we calculate their learned profiles’ cosine distance, and predict them to be friends if the cosine distance is below a chosen threshold.¹³ Note that our prediction does not need the knowledge of any existing friendships, therefore, it is unsupervised.

5.2 Evaluation

Experimental Setup. For each dataset, we randomly sample the same number of stranger pairs as the number of friend pairs to construct the negative cases. Then, we compute cosine distances for all friend and stranger pairs. AUC (area under the ROC curve) is adopted as the evaluation metric [3, 18, 56, 57]. There are mainly three hyperparameters in our model, i.e., walk length, walk times, and the dimension of the learned profile. Based on cross validation, we set them to 120, 80, and 512, respectively.

Baseline Models. We further establish three baseline models following the traditional link prediction approach [29], namely Jaccard

¹²Following previous works [3, 17], $N(v)$ includes 10 nodes precedent and after v in all the random walk traces.

¹³We also learn a set of vectors for all hashtags, as we concentrate on friendship prediction, these hashtag vectors are simply neglected.

Table 7: Formal definition of baseline models for friendship prediction. $\mathcal{H}(u)$ is the set containing all the hashtags u has shared.

Baseline	Definition
Common hashtags	$ \mathcal{H}(u) \cap \mathcal{H}(u') $
Jaccard index	$\frac{ \mathcal{H}(u) \cap \mathcal{H}(u') }{ \mathcal{H}(u) \cup \mathcal{H}(u') }$
Preferential attachment	$ \mathcal{H}(u) \cdot \mathcal{H}(u') $

index, common hashtags, and preferential attachment. Their formal definition is presented in Table 7.

Results. Figure 8 depicts the AUCs for friendship prediction. In all the three datasets, we obtain more than 0.8 AUC which shows that hashtags can provide strong signals on inferring social relations. In particular, the prediction in the New York dataset achieves the best performance.

Our approach also outperforms all the three baseline models significantly. For the best baseline model, i.e., Jaccard index, we achieve around 20% performance gain in all the datasets. One of the major advantages of our approach is its ability to predict two users’ friendship regardless of whether they share common hashtags. We observe that even for pairs of users sharing no common hashtags, our prediction still achieves a decent performance in all the datasets: 0.759 AUC in New York, 0.733 AUC in Los Angeles, and 0.728 AUC in London. This further demonstrates the effectiveness of our approach.

We also study the influence of hyperparameters on the prediction performance. Figure 9 depicts the heatmap when jointly tuning walk times and walk length. We see that when these two values are small, i.e., 40, the prediction results are relatively weak in all the three datasets. When increasing both, the performance gets better. However, the relation between the prediction result and the magnitude of the hyperparameters is not monotonic: When we set both walk times and walk length to 200, the prediction results drop in all cases. Note that we also perform the same study on the dimension of each learned profile, and observe that 512 leads to the best prediction.

6 RELATED WORK

Hashtags provide us with an unprecedented chance to understand the modern society. Researchers have studied hashtags from many perspectives [2, 5, 6, 8–10, 12, 14–16, 23, 24, 26–28, 30, 33, 36, 37, 41–43, 49, 52, 54, 55, 58, 60, 63, 65].

Souza et al. use #selfie to study the phenomenal self-portrait behavior in OSNs [58]. They collect a large dataset from Instagram, and show that the amount of posts associated with #selfie increases 900 times from 2012 to 2014. Moreover, their results suggest that posts with #selfie attract more likes and comments than others. The authors further show that there exist cultural variations of selfie behavior across countries.

Mejova et al. study food-related hashtags, such as #foodporn, on Instagram to understand people’s dining behavior on a global scale [37]. They first show that desserts in social media are dominating over local cuisines. Then, they discover through hashtags

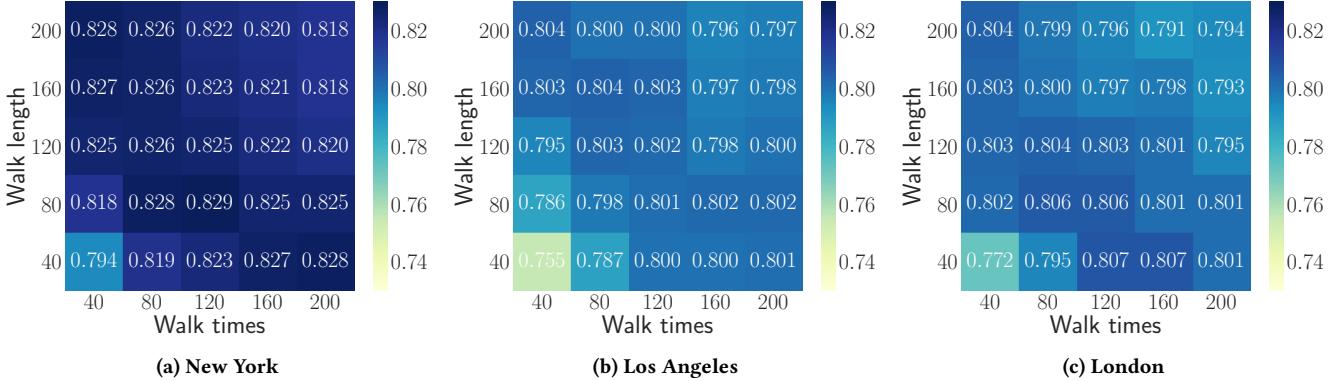


Figure 9: Influence of walk length and walk times on the prediction performance in three datasets. Each value in the matrix represents the corresponding AUC.

that food can motivate people to engage in a healthier life style. Similar to #selfie, Mejova et al. show that posts associated with healthy-related hashtags attract more likes than others.

David et al. use hashtags together with the smiley face emoji to perform sentimental analysis on Twitter [10]. In this work, hashtags are used as the sentimental labels, and the authors summarize features over tweets. Evaluation results show that their classification achieves very effective performance.

Multiple works also use hashtags to study various political movements. Olteanu et al. analyze the demographics behind #blacklivesmatter, and conclude that African-Americans and young females engage more with the hashtag than others [42]. Manikonda et al. perform a comparative analysis of #metoo shared on Twitter and Reddit [34]. They observe that posts on Reddit concentrate on sharing personal stories while tweets express public support for the #metoo movement.

Zhang et al. study the privacy implications of sharing hashtags [65]. They first utilize a simple bag-of-words model and a random forest classifier to perform location prediction based on hashtags. Evaluation shows that their approach achieves more than 70% accuracy over fine-grained locations. Then, the authors propose a privacy-preserving system, namely Tagvisor, to mitigate the location privacy risks. Tagvisor implements three different obfuscation mechanisms including hiding, replacement, and (location category) generalization. Extensive experiments suggest that by obfuscating two hashtags, Tagvisor can successfully mislead the location prediction model, while maintaining a high-level utility with respect to hashtag semantics.

While the above works concentrate on either a certain type of hashtags or a certain property of hashtags, Ferragina et al. perform a general analysis of hashtags on Twitter [12]. In particular, they concentrate on the semantics of Twitter's hashtags. They first build a hashtag-entity graph over a large number of tweets. Then, the authors perform two natural language processing tasks, namely hashtag relatedness and hashtag classification, based on the features extracted from the constructed hashtag-entity graph. Experimental results show that this approach outperforms state-of-the-art solutions by a large extent. Besides targeting a different OSN than

Ferragina et al. [12], we perform a much broader analysis on hashtags with several new angles, such as spatial pattern, semantic displacement, and social signal.

7 CONCLUSION

In this paper, we perform the first large-scale analysis on understanding hashtags shared on Instagram. Our study is centered around three research questions which aim at understanding hashtags from three different dimensions, i.e., the temporal-spatial dimension, the semantic dimension, and the social dimension. We collect three large datasets from Instagram containing more than 7 million hashtags shared over 5 years to perform our analyses.

We first show that hashtags can be categorized into four different clusters according to their temporal patterns, and people are more willing to share hashtags at certain places, such as parks. We then discover that some hashtags indeed exhibit large semantic displacement. Moreover, we propose a notion namely hashtag entropy and show the strong negative correlation between hashtag entropy and semantic displacement. In the end, we propose a bipartite graph embedding model to summarize users' hashtag profiles and rely on these profiles to perform friendship prediction. The effective prediction performance suggests that there is a strong connection between hashtags and social relations.

8 ACKNOWLEDGMENTS

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0656). The author would like to thank Dr. Sandra Strohbach and Dr. Mathias Humbert for their constructive feedback on the manuscript.

REFERENCES

- [1] Luca Maria Aiello and Nicola Barbieri. 2017. Evolution of Ego-networks in Social Media with Link Recommendations. In *Proceedings of the 2017 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 111–120.
- [2] Jisun An and Ingmar Weber. 2016. #greysanatomy versus #yankees: Demographics and Hashtag Use on Twitter. In *Proceedings of the 2016 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 523–526.
- [3] Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. 2017. walk2friends: Inferring Social Links from Mobility Profiles. In *Proceedings of the 2017 ACM*

- SIGSAC Conference on Computer and Communications Security (CCS). ACM, 1943–1957.
- [4] Lars Backstrom and Jure Leskovec. 2011. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *Proceedings of the 2011 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 635–644.
 - [5] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *CoRR abs/1802.04291* (2018).
 - [6] Yi Chang, Jiliang Tang, Dawei Yin, Makoto Yamada, and Yan Liu. 2016. Timeline Summarization from Social Media with Life Cycle Models. In *Proceedings of the 2016 International Joint Conferences on Artificial Intelligence (IJCAI)*. IJCAI, 3698–3704.
 - [7] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proceedings of the 2011 ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 1082–1090.
 - [8] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. 2011. Political Polarization on Twitter. In *Proceedings of the 2011 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 89–96.
 - [9] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. 2012. Discover Breaking Events with Popular Hashtags in Twitter. In *Proceedings of the 2012 ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 1794–1798.
 - [10] Dmitry David, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smilies. In *Proceedings of the 2010 International Conference on Computational Linguistics (Coling)*. ACL, 241–249.
 - [11] Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nittiest Chawla, Jinghai Rao, and Huanhuan Cao. 2012. Link Prediction and Recommendation across Heterogeneous Social Networks. In *Proceedings of the 2012 International Conference on Data Mining (ICDM)*. IEEE, 181–190.
 - [12] Paolo Ferragina, Francesco Piccinno, and Roberto Santoro. 2015. On Analyzing Hashtags in Twitter. In *Proceedings of the 2015 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 110–119.
 - [13] Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. 2018. BiNE: Bipartite Network Embedding. In *Proceedings of the 2018 International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 715–724.
 - [14] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In *Proceedings of the 2016 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 33–42.
 - [15] Ruth Garcia Gavilanes, Neil O'Hare, Luca Maria Aiello, and Alejandro Jaimes. 2013. Follow My Friends This Friday! An Analysis of Human-Generated Friendship Recommendations. In *Proceedings of the 2013 International Conference on Social Informatics (SocInfo)*. Springer, 46–59.
 - [16] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. 2013. Data Portraits: Connecting People of Opposing Views. *CoRR abs/1311.4658* (2013).
 - [17] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 2016 ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 855–864.
 - [18] Inken Hagedstedt, Yang Zhang, Mathias Humbert, Pascal Berrang, Haixu Tang, Xiaofeng Wang, and Michael Backes. 2019. MBeacon: Privacy-Preserving Beacons for DNA Methylation Data. In *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS)*. Internet Society.
 - [19] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 1489–1501.
 - [20] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the 2017 Annual Conference on Neural Information Processing Systems (NIPS)*. NIPS, 1025–1035.
 - [21] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. In *Proceedings of the 2014 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 595–598.
 - [22] Xiao Huang, Qingquan Song, Jundong Li, and Xia Hu. 2018. Exploring Expert Cognition for Attributed Network Embedding. In *Proceedings of the 2018 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 270–278.
 - [23] Jin Yea Jang, Kyungskik Han, Patrick C. Shih, and Dongwon Lee. 2015. Generation Like: Comparative Characteristics in Instagram. In *Proceedings of the 2015 Annual ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, 4039–4042.
 - [24] Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, and Zhe Zhao. 2014. Predicting Bursts and Popularity of Hashtags in Real-Time. In *Proceedings of the 2014 International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 927–930.
 - [25] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 2015 International Conference on World Wide Web (WWW)*. ACM, 625–635.
 - [26] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. 2012. On Recommending Hashtags in Twitter Networks. In *Proceedings of the 2012 International Conference on Social Informatics (SocInfo)*. Springer, 337–350.
 - [27] Janette Lehmann, Bruno Goncalves, José J. Ramasco, and Ciro Cattuto. 2012. Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 2012 International Conference on World Wide Web (WWW)*. ACM, 251–260.
 - [28] Cheng-Te Li, Chien-Yuan Wang, Chien-Lin Tseng, and Shou-De Lin. 2011. MemTube: A Sentiment-based Audiovisual System for Analyzing and Displaying Microblog Messages. In *Proceedings of the 2011 Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 32–37.
 - [29] David Liben-Nowell and Jon Kleinberg. 2007. The Link-prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology* (2007).
 - [30] Yu-Ru Lin, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. 2013. #Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtags. In *Proceedings of the 2013 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 370–379.
 - [31] Qingyun Liu, Shiliang Tang, Xinyi Zhang, Xiaohan Zhao, Ben Y. Zhao, and Haitao Zheng. 2016. Network Growth and Link Prediction Through an Empirical Lens. In *Proceedings of the 2016 ACM Internet Measurement Conference (IMC)*. ACM, 1–15.
 - [32] Yao Ma, Zhaochun Ren, Ziheng Jiang, Jiliang Tang, and Dawei Yin. 2018. Multi-Dimensional Network Embedding with Hierarchical Structure. In *Proceedings of the 2018 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 387–395.
 - [33] Zongyang Ma, Aixin Sun, and Gao Cong. 2012. Will This #Hashtag Be Popular Tomorrow?. In *Proceedings of the 2012 International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 1173–1174.
 - [34] Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. 2018. Twitter for Sparking a Movement, Reddit for Sharing the Moment: #metoo through the Lens of Social Media. *CoRR abs/1803.08022* (2018).
 - [35] Lydia Manikonda, Yuheng Hu, and Subbarao Kambhampati. 2014. Analyzing User Activities, Demographics, Social Network Structure and User-Generated Content on Instagram. *CoRR abs/1410.8099* (2014).
 - [36] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding Semantics to Microblog Posts. In *Proceedings of the 2012 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 563–572.
 - [37] Yelena Mejova, Sofiane Abbar, and Hamed Haddadi. 2016. Fetishizing Food in Digital Age: foodporn Around the World. In *Proceedings of the 2016 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 250–258.
 - [38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 2013 International Conference on Learning Representations (ICLR)*.
 - [39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositional Properties. In *Proceedings of the 2013 Annual Conference on Neural Information Processing Systems (NIPS)*. NIPS, 3111–3119.
 - [40] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the 2013 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 400–408.
 - [41] Ferda Ofli, Yusuf Aytar, Ingmar Weber, Raggi al Hammouri, and Antonio Torralba. 2017. Is Saki #delicious?: The Food Perception Gap on Instagram and Its Relation to Health. In *Proceedings of the 2017 International Conference on World Wide Web (WWW)*. ACM, 509–518.
 - [42] Alexandra Olteanu, Ingmar Weber, and Daniel Gatica-Perez. 2016. Characterizing the Demographics Behind the #BlackLivesMatter Movement. *CoRR abs/1512.05671* (2016).
 - [43] Raphael Ottoni, Diego Las Casas, Joao Paulo Pesce, Wagner Meira Jr, Christo Wilson, Alan Mislove, and Virgilio Almeida. 2014. Of Pins and Tweets: Investigating How Users Behave Across Image- and Text-Based Social Networks. In *Proceedings of the 2014 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 386–395.
 - [44] Jun Pang and Yang Zhang. 2015. Location Prediction: Communities Speak Louder than Friends. In *Proceedings of the 2015 ACM Conference on Online Social Networks (COSN)*. ACM, 161–171.
 - [45] Jun Pang and Yang Zhang. 2017. Quantifying Location Sociability. In *Proceedings of the 2017 ACM Conference on Hypertext and Social Media (HT)*. ACM, 145–154.
 - [46] Beatrice Perez, Mirco Musolesi, and Gianluca Stringhini. 2018. You Are Your Metadata: Identification and Obfuscation of Social Media Users using Metadata Information. In *Proceedings of the 2018 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 241–250.
 - [47] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 2014 ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 701–710.
 - [48] Huy Pham, Cyrus Shahabi, and Yan Liu. 2013. EBM: An Entropy-Based Model to Infer Social Strength from Spatiotemporal Data. In *Proceedings of the 2013*

- ACM SIGMOD International Conference on Management of Data (SIGMOD).* ACM, 265–276.
- [49] Lisa Posch, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2013. Meaning as Collective Use: Predicting Semantic Hashtag Categories on Twitter. In *Proceedings of the 2013 International Conference on World Wide Web (WWW)*. ACM, 621–628.
- [50] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In *Proceedings of the 2018 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 459–467.
- [51] Meng Qu, Jian Tang, and Jiawei Han. 2018. Curriculum Learning for Heterogeneous Star Network Embedding via Deep Reinforcement Learning. In *Proceedings of the 2018 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 468–476.
- [52] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. 2014. The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City. In *Proceedings of the 2014 ACM Conference on Hypertext and Social Media (HT)*. ACM, 116–125.
- [53] Suhas Ranganath, Ghazaleh Beigi, and Huan Liu. 2018. Curriculum Learning for Heterogeneous Star Network Embedding via Deep Reinforcement Learning. In *Proceedings of the 2018 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 468–476.
- [54] Jacob Ratkiewicz, Michael D. Conover, Mark Meiss, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and Tracking Political Abuse in Social Media. In *Proceedings of the 2011 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 297–304.
- [55] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *Proceedings of the 2011 International Conference on World Wide Web (WWW)*. ACM, 695–704.
- [56] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS)*. Internet Society.
- [57] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. 2011. Exploiting Place Features in Link Prediction on Location-based Social Networks. In *Proceedings of the 2011 ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 1046–1054.
- [58] Flavio Souza, Diego de Las Casas, Vinicius Flores, SunBum Youn, Meeyoung Cha, Daniele Quercia, and Virgilio Almeida. 2015. Dawn of the Selfie Era: The Whos, Wheres, and Hows of Selfies on Instagram. In *Proceedings of the 2015 ACM Conference on Online Social Networks (COSN)*. ACM, 221–231.
- [59] Jian Tang, Meng Qu, Mingzhu Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 2015 International Conference on World Wide Web (WWW)*. ACM, 1067–1077.
- [60] Oren Tsur and Ari Rappoport. 2012. What's in a Hashtag? Content based Prediction of the Spread of Ideas in Microblogging Communities. In *Proceedings of the 2012 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 643–652.
- [61] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. 2009. On the Evolution of User Interaction in Facebook. In *Proceedings of the 2009 ACM Workshop on Online social networks (WOSN)*. ACM, 37–42.
- [62] Hongwei Wang, Fuzheng Zhang, Min Hou, Xing Xie, Minyi Guo, and Qi Liu. 2018. SHINE: Signed Heterogeneous Information Network Embedding for Sentiment Link Prediction. In *Proceedings of the 2018 ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 592–600.
- [63] Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We Know What You #Tag: Does the Dual Role Affect Hashtag Adoption?. In *Proceedings of the 2012 International Conference on World Wide Web (WWW)*. ACM, 261–270.
- [64] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Hierarchical Graph Representation Learning with Differentiable Pooling. *CoRR abs/1806.08804* (2018).
- [65] Yang Zhang, Mathias Humbert, Tahleen Rahman, Cheng-Te Li, Jun Pang, and Michael Backes. 2018. Tagvisor: A Privacy Advisor for Sharing Hashtags. In *Proceedings of the 2018 Web Conference (WWW)*. ACM, 287–296.