Measuring and Summarizing Movement in Microblog Postings

Eduardo Ruiz[*] and **Vagelis Hristidis**
University of California-Riverside
{eruiz009,vagelis}@cs.ucr.edu

Carlos Castillo[†]
Qatar Computing Research Institute
chato@acm.org

Aristides Gionis[‡]
Aalto University
aristides.gionis@aalto.fi

Abstract

Every day, users publish hundreds of millions of microblog postings in popular social-networking platforms such as Twitter and Facebook. When considered in aggregation, microblog postings have been shown to exhibit temporal patterns that reflect events of global significance.

In this paper, we propose techniques to identify and quantify spatial patterns: for instance, a hashtag that is popular in one city on a given day, may become popular in a different city on the next day. Detecting these patterns is challenging given that the data are noisy and posts are not physically moving, i.e., they are not continuous trajectories in space like vehicles. Second, we introduce a multi-granular summarization model to describe the movement of a hashtag between two time periods. For interpretability, we seek representations of spatial changes that follow natural or administrative boundaries on a map, such as cities and states.

We compare various movement measures using quantitative approaches and user surveys. We evaluate our movement summarization schemes by analytical loss and coverage functions. Our results show that it is possible to reliably detect relevant spatial changes automatically, and to produce simple summaries that represent accurately these changes.

1 Introduction

The development of methods for mining the microblog posting patterns of users has attracted a significant amount of research in recent years. Some of these studies have focused on modeling temporary "bursts" of activity about a topic (Mathioudakis, Bansal, and Koudas 2010) including "bursts" around a specific location (Lappas et al. 2012), or on connecting changes in the activity of a topic to other time series in the epidemic (Ginsberg et al. 2009) and financial (Ruiz et al. 2012) domains, among many others.

In contrast, modeling spatial changes in aggregate microblog activity – we focus on activities represented by a hashtag in this paper, although the same techniques can be applied to other representations – has not been studied at the same depth. There is a large body of work covering the topic of mining movements and trajectories of mobile objects such as cars (e.g., (Vieira, Bakalov, and Tsotras 2010)) or mobile phones (e.g., (González, Hidalgo, and Barabási 2008; Hazas, Scott, and Krumm 2004)). However, these problems are fundamentally different from the problem we study. In our framework, since a microblog hashtag does not "move", we do not study trajectories recording the movement of specific objects; instead we study geographical movement of aggregated actions recorded over millions of users.

The first question we study is how to measure changes on the geo-spatial distribution of a hashtag between two time intervals, that is, how to obtain the hashtags that are moving in space and how to quantify such movement. To the best of our knowledge, there are no proposed metrics for this problem. In the remainder of this paper we refer to this phenomenon as *movement*, but we remark that often these location changes do not constitute a continuous movement in physical space, but discrete steps between days.

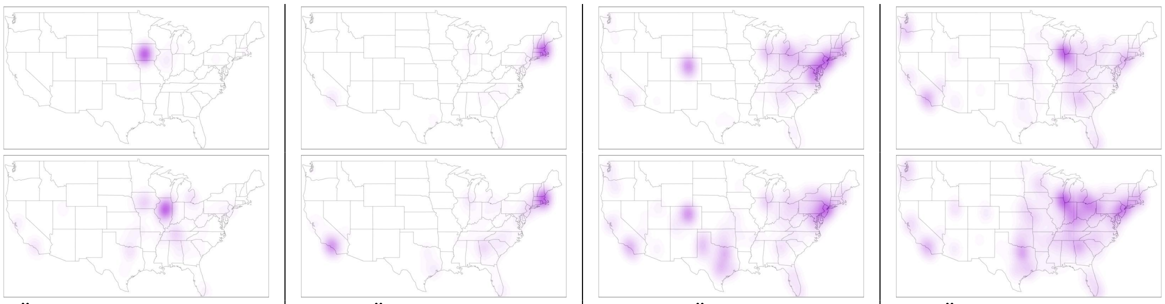
In our experiments, we use a dataset from Twitter and select a set of hashtags. We use user-provided coordinates plus a set of simple heuristics to find the locations on a subset of those postings. Given that, our main focus is to model hashtag movements and develop movement-detection algorithms. Our results are independent from the specific dataset, and from the tweet-selection and geo-location methods used. Hence, topic detection algorithms can be used instead of hashtags, and more sophisticated location estimation methods can be employed as discussed below.

Figure 1 shows example time frames for different hashtags. We observe the diversity of changes that can be defined as a movement. The simplest one is a linear movement from point A to point B, like in the case of **`#LadyAntebellum`**, which refers to a small music band that is on tour from one city to another. Sometimes there can be multiple centers, such as basketball team **`#Celtics`** that splits between their home city and the city they are visiting (or multiple cities depending on the importance of the game). Other movements like a **`#Snow`** storm are not point to point, but move across multiple regions. Finally, a spread from one area to multiple areas can also be considered movement, as in the case of the interest on rock band **`#Nickelback`**. These examples serve to illustrate some of the challenges in measuring hashtag movements, which include:

[*] Most of this work was done while the author was a summer intern in Yahoo! Research Barcelona.

[†] Work done while at Yahoo! Research.

[‡] Work done while at Yahoo! Research.



#LadyAntebellum (point-wise movement)

#Celtics (split)

#Snow (regional movement)

#Nickelback (expansion)

Figure 1: Example of hashtags that move between two consecutive days (above: first day, below: second day). Darker areas have a larger number of messages on the hashtag. We have included the corresponding hashtag and a qualitative description (in parentheses) of the type of movement observed.

- There are no well-defined start and end points, which would allow to use simple linear distance measurements.
- Many different types of spatial changes can be observed, from simple point-to-point movement, to concentration around a specific location, or expansion from one to multiple locations, etc.
- The volume of activity may change dramatically between the two observation periods.
- The movement measure must agree with what users visually judge as moving.

Problem 1 [Measure Movement]: Our first objective is to quantify the geo-spatial change between two time frames. This measure can be used to detect interesting events for a particular hashtag. In the same way that shifts in the language model of posts of a hashtag can be considered as event frontiers, we also believe that having a measure of movement can be used as an interesting feature for finding relevant milestones. Detecting geographical movement of hashtags (or of a topic if a topic classifier is used) can have many applications, e.g., adapt marketing campaigns or product supplies (when people talk about a local product in other cities), or detect the spread of ideas (Conover MD 2013), trends, or diseases. We can also use the movement measure to detect outlier changes for mainly local hashtags. We can view the movement amount of a hashtag in consecutive days as a time series, and detect unexpected behavior. Finally, abnormal movements can be used as a feature to detect spam messages unrelated to a hashtag.

Problem 2 [Summarize Movement]: Our second objective is to develop methods that summarize movement between two time periods in a compact, informative and intuitive way. In our experiments, we focus on methods that generate summaries at the level of well-understood administrative regions such as cities, counties, states, etc. Consider for instance the US Center for Disease Control (CDC) that monitors microblog postings containing the word "*cholera*." If an outburst is detected and the CDC must consult with a relatively focused set of local public health agencies, which ones should be selected? Intuitively, the locations with high activity or high change of activity are of interest.

The overall challenge of any (lossy) summarization procedure is the trade-off between accuracy and conciseness. We define a *loss function* to measure the accuracy loss. A key challenge is that in the summary we should represent areas with high activity, as well as areas with a large amount of change, from the multi-level spatial hierarchy. For instance, a hashtag may spread from a city to a whole state.

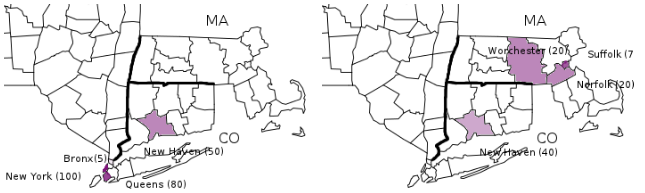
Summary of contributions:

- We introduce and model the problem of geospatial movement of microblogging activity (Section 2).
- We propose and compare three classes of approaches to quantify the microblogging movement. We conduct user studies to identify the measure that is most correlated with what users consider as movement (Section 3).
- We study the problem of summarizing the movement of microblogging activity. We propose multi-granular approaches that account for both the high activity and high movement locations in a principled way. We present an efficient algorithm for the movement summarization problem (Section 4).
- We conduct thorough performance experiments and user studies on Twitter data that show that our techniques are both efficient and intuitive (Section 5).

We present related work in Section 6 and we offer our concluding remarks in Section 7.

2 Framework and Problem Statement

Our general problem can be informally described as follows: given a hashtag that has been of interest to Internet users over time, describe changes in the position of the users who use the hashtag.In this work we use Twitter as a data source, due to its popularity and the possibility of accessing public data from millions of users. However, our work can be applied to any setting where posts can be classified to topics and have spatiotemporal characteristics (location and timestamp).When a tweet does not have spatial coordinates we use the location field of the author user as a proxy. In our experiment we ignore posts with neither post nor user coordinates.



(a) Day 1

(b) Day 2

Figure 2: Example depiction of a hashtag "moving" from New York City to the state of Massachusetts. A darker shade indicates more posts. Thin lines are county boundaries, thick lines are state boundaries.

There are a number of methods for determining topics (e.g., a famous person, or a real-world event) in microblogging postings. We use the simplest of them which is to look at the hashtag information already provided by the users through hashtags (starting with #). More elaborate methods for topic and location detection (see Section 6) would not change the design of our framework or algorithms.

We define a *frame* F as the distribution of the locations of all the messages about a hashtag that appear in a particular interval of time. The name "frame" is an analogy to video technology where each static image forming a video captures the current state of the object being recorded on a particular moment. For instance, we can define a single-day frame for **`#Lakers`** from 2012-03-24 00:00:00 to 2012-03-24 23:59:59 which contains the locations of all posts that contain the hashtag **`#Lakers`** during this timeframe.

Our objective is to automatically detect what users would visually consider "movement". Given two frames, movement is the change in the positions of the posts between the two frames. Ideally, we want to establish a real value that can be used to quantify movement. We define the movement-quantification problem as follows:

Problem 2.1 (Measure Movement) *Given two frames (of the same hashtag), quantify the geospatial distance between them.*

A second problem is how to concisely present the information about the movement to the user. This representation can be used to describe similar movements in different days or to find common patterns of movement for a particular hashtag or set of hashtags.

Problem 2.2 (Summarize Movement) *Given two frames, for which movement has been established, describe the movement in a succinct way.*

For instance, Figure 2 shows two frames in different days for a particular hashtag that *moves* from the state of New York to the state of Massachusetts. The color shades show the level of activity (number of postings) in each particular region in each of the two time periods. We can visually see a shift from a certain amount of activity in New York City on the first day, to a lower level of activity in Boston on the second day.

In this example, the following statement can be considered a good summarization of the hashtag movement: "the hashtag moved from New York City on Monday, to the State of Massachusetts on Tuesday, and the overall activity decreased on the second day."

3 Movement Measures

We discuss three general classes of models for the movement measurement problem: (1) Single-center models: given two frames we find a single center for each frame, and we compare the centers. (2) Multiple-center models: given two frames we measure the movement as the amount of changes that are necessary to make both frames equal. (3) Similarity models: given two frames we measure the movement as the number of elements that remain in a similar place. This definition essentially measures how similar are the frames.

3.1 Single-Center Models

Centroid distance: In this model, the activity in each frame is summarized by a single centroid, which is the centroid of all posting locations in this frame. We calculate the distance between these centroids using Euclidean distance.

The centroid model is intuitive when there is a single source in the initial frame (city, state) and a single destination on the second frame. However in many cases we find multiple sources/destinations or outliers that influence the centroid location, in the sense that the centroid is not an accurate representation of the global micro-blogging activity. Despite those conceptual shortcomings, this simple model has a decent performance in practice, as we show in Section 5.

Hotelling measure: This model is based on the natural question of how to decide if the change of the centroid location is really relevant, and has not occurred simply by chance. To decide if the movement is relevant we apply a statistical significance test against the hypothesis that both frames actually depict the same geographical distribution.

To apply our test, we first need to make some assumptions on the nature of the distribution. We define $g(t)$ as the geolocation of tweet t given by (latitude, longitude) pairs. We consider that the tweets in both frames are generated from a multi-variate normal distribution with mean μ and covariance Σ

$$p(g(t)) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-(g(t)-\mu)\Sigma^{-1}(g(t)-\mu)}$$

where both parameters μ and Σ can be estimated using the simple maximum-likelihood estimators for each frame.

We consider the Hotelling test, a generalization of the well-known t-test, to decide if the movement is significant. The test assumes that the centroids of the two frames are the same, and uses the data to reject this hypothesis. To simplify our presentation we describe this as a two-step process. First we calculate the function:

$$t^2 = \frac{n_1 \times n_2}{n_1 + n_2}(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

where $\hat{\mu_1}, \hat{\mu_2}$ are the centroids for each frame, and $\hat{\Sigma}$ is a weighted sum of the individual covariances of each frame. Notice that the value of t increases as we have a bigger difference between frames and can be reduced depending on the spreading explained by the covariance.

After we calculate the estimator t^2 we compare against the Fisher distribution to decide if we should reject our default hypothesis. Given a confidence α we reject the null hypothesis if:

$$\frac{n_1 + n_2 - 3}{2(n_1 + n_2 - 2)} t^2 \geq F_{2, n_1+n_2-3, \alpha}$$

where n_1, n_2 are the numbers of tweets on each frame. Note that the test is affected by the number of samples we have in each day. We rank the movements using the minimum α we can use before we fail to reject the null hypothesis.

3.2 Multiple-Center Models

Earth-mover distance (EMD measure): A limitation of single-center models is that they assume the presence of one center of activity in each frame. Backstrom et al. (2008), in the context of geo-location of web-search queries, propose a solution where actions are first clustered and then several centers are learned. Further, we can consider a model where centers are detected using a mixture of multiple Gaussians (Bishop 2006). Two major challenges for dealing with multiple centers are the following: (i) decide what is the appropriate number of centers for each frame; and (ii) decide how the centers map to each other on the movement.

To overcome both of these issues we employ a non-parametric model of the distribution of posts. To introduce our model we first need to define the following terms: Given a geospatial region R, a *region partitioning* (or simply partitioning) is a set $r_1 \ldots r_s$ of *cells* of R that partition R. We do not restrict the elements to be uniform in size or shape. For example, we can use square cells that have a particular area, or the administrative geographical division of the region, e.g., states or counties. Figure 2(a) depicts an example of a partition of the North-East US region R using county cells.

Given a partitioning of R, we can assign a specific number to each cell. This can be the number of tweets in a particular frame.

Definition 3.1 *1 [Frame Partitioning] Given a partitioning $r_1 \ldots r_s$ of a region and a frame F, the frame partitioning f of F is a function that assigns a weight w_j to each cell r_j, i.e., $f(r_j) = w_j$.*

Figure 2(a) shows the frame partitioning of a hashtag for the North-East US region. We highlight the cells that have some activity. We also show the name of the cells (counties) and the level of activity (number of postings) in parentheses.

We define the distance between two frame partitionings by adapting the Earth-mover distance (EMD) (Rubner, Tomasi, and Guibas 2000) to our problem. Informally, this represents the amount of change that would be required to make the two frames equal.

Given two frame partitioning f and g for region partitioning $r_1 \ldots r_s$, we define the $EMD(f,g)$ as:

$$EMD(f,g) = \min \sum_{i=1}^{s} \sum_{j=1}^{s} w_{ij} d(r_i, r_j)$$

constrained by

$$\begin{aligned}
\sum_{j=1}^{s} w_{i,j} &\leq f(r_i), \\
\sum_{i=1}^{s} w_{i,j} &\leq g(r_j), \text{ and} \\
\sum_{i=1}^{s} \sum_{j=1}^{s} w_{i,j} &= \min(\sum_{i=1}^{s} f(r_i), \sum_{j=1}^{s} f(r_j))
\end{aligned}$$

The components w_{ij} in the optimal solution, can be interpreted as the amount of mass (number of posts) that is exchanged from r_i on the first frame to the region r_j on the second frame.

The function $d(r_i, r_j)$ measures the distance between two cells on the partition. We use the euclidean distance between the centers of both cells.

EMD$_{\text{log}}$ measure: We may want to decrease the effect of distance in the above described EMD measure. For instance, we may view a movement from New York to Los Angeles only slightly more important than one movement from New York to Boston. We found that the latter interpretation is intuitive when evaluating Twitter movements. We refer to the EMD variant where we take the logarithm of the distance as EMD$_{\text{log}}$.

3.3 Similarity-Based Models

The intuition behind similarity-based models is that frames with movement would have a low overlap between cells. That is, we measure the amount of change in each cell, without considering their actual distances.

Cosine measure: Zhang (Zhang et al. 2012) utilize the cosine similarity between partitions as a way to cluster correlated tags in the geospatial domain. We consider a modification of the cosine measure to measure the difference between two frames. Using the partitioning notation introduced in Section 3.2, let $f(R)$ be the frame partitioning vector for f, that is, $f(R) = (f(r_1), \ldots, f(r_s))$. Then, we can write the cosine similarity as:

$$\cos(f,g) = \frac{f(R)\, g(R)}{||f(R)||\, ||g(R)||}$$

CK measure: We explore techniques that consider the compression ratio between two frames as a measure of similarity (Li et al. 2003). In particular, we consider the work of Campana et al. (2010) that describe a compression measure based on video compression. Given two frames f, g we create a video that only has these two frames. Let $\text{mpegSize}(f,g)$ be the size of the compressed video obtained using a common MPEG-1 encoding. Then the CK measure is:

$$\text{CK}(f,g) = \frac{\text{mpegSize}(f,g) + \text{mpegSize}(g,f)}{\text{mpegSize}(f,f) + \text{mpegSize}(g,g)} - 1$$

To create the images for each frame f, g we use the following process: first we take the frame partitioning and create a base histogram with a value for each cell. Then we smooth the value of each cell using the kernel density-estimation method. The smoothed histogram is used to build a heat map where darker areas mean higher activity.

Table 1 summarizes all the proposed movement measures.

4 Movement summarization

To motivate our summarization strategy, consider again the movement between frames in Figures 2(a) and 2(b) for a particular hashtag. One possible way to represent this movement is to list all the counties that are active on both frames.

Table 1: Movement measures

Measure	Description	Comments
Centroid	Centroids Distance	Distance and direction.
Hotelling	Hotelling significance Rest	Sample size.
EMD	Earth Mover Distance	Distance/Multiple centers.
$\mathrm{EMD}_{\log}$	Dampening EMD,	Distance/Multiple centers.
Cosine	Cosine Histograms Vectors	Overlap / Ignore Distance.
CK	Compression ratio of MPEG	Overlap/ Ignore Distance.

In our example, this list would have nine counties (four sources and five destinations). This representation can be quite verbose. Furthermore, the representation is not succinct and it does not help the user to understand the movement rapidly.

To address the points above, we propose a two-part summarization approach. The first part summarizes the original state before the movement, that is, the first frame. The second part of the summary represents the changes that are necessary to approximately re-create the second frame from the first. This idea of delta-frames has been extensively used for video compression (Nasrabadi N.M. Lin 1989). To reduce the size of both parts we use the hierarchical decomposition of the regions to summarize multiple cells with a single super-cell.

In our example, the movement could be expressed now in the following way:

Base-frame summary : New_York.NY, Queens.NY
New_Haven.CO.
δ-frame summary : (-) New_York.NY, (-) Queens.NY,
(+) MA.

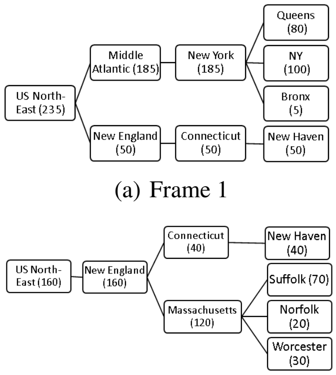
The first part of the summary takes the three most important areas from the original frame and it ignores a county with low activity (Bronx.NY). The delta summary is used to reconstruct the movement. In this case the state of Massachusetts becomes more active and the counties on New York become dormant. The New Heaven County is not represented in the second summary, so we assume that the activity is unchanged there. Using the state instead of all the counties in Massachusetts we reduce the size of the summary, balancing size with informativeness.

To represent the containment between regions we use a hierarchical representation of the region R.

Definition 4.1 (Partition Tree) *A partition tree Ψ groups the partition cells of R using an h-level hierarchy. Each leaf corresponds to one cell in the partition. An internal node (super-cell) n is the union of all its descendants. Cells on Ψ are augmented with scores that correspond to the particular activity on the corresponding leaf.*

We use $R(n)$ to refer to the area covered by a super-cell n. We say that cell r_j is *covered* by a super-cell n, if the area $R(r_j)$ is included in the area $R(n)$. For the rest of the work the hierarchy is based on the administrative division of the country. Cells are the leaves of the tree (counties) and super-cells (internal nodes) are states or regions.

We now define the equivalent of the frame partitioning with respect to a partition tree, which assigns counts to each cell of the partition tree:



(a) Frame 1

(b) Frame 2

Figure 3: F-Tree for administrative hierarchy

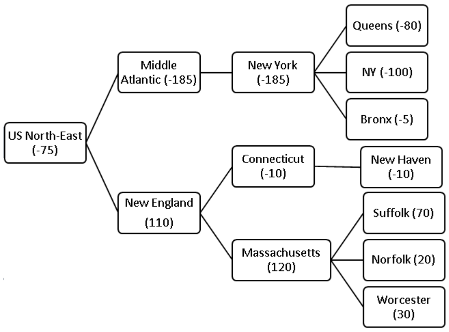


Figure 4: Differential tree for geographical hierarchy

Definition 4.2 (Frame Partition Tree) *Given a partition tree Ψ, the F-Tree for frame F, is a function that assigns a score (count) to each node of Ψ. The score of a leaf node is the same as in the frame partitioning of Ψ. The score of an internal node ("super-cell") is the sum of the scores of all its descendant leaves.*

In Figure 3 we show two F-Trees where each leaf is annotated with its activity (number of posts). We do not display cells or super-cells with zero values in this example, but these are also part of the tree.

To represent the movement between two frames we combine their F-Trees, to create a representation of the differential. We define this difference using the activity change between frames on each cell of the F-Trees. More formally, we have the following definition:

Definition 4.3 (Differential F-Tree) *Given two F-Trees F_1 and F_2 the Differential F-Tree (δ-Tree) is a function δ that assigns the score δ-Tree$(n) = F_1(n) - F_2(n)$, to each cell $n \in \Psi$.*

Figure 4 shows the δ-Tree for the two trees presented in Figure 3. A δ-Tree cell has negative score if the original frame has fewer posts than the destination frame for this cell. Also note that the cell scores of an F-Tree are monotone (increasing) as we go higher on the tree, which is not the case for δ-Tree.

The movement summarization problem can then be defined as follows:

Problem 4.4 (Movement Summarization) *Given the F-Tree f of the original frame, the δ-Tree δ between the orig-*

inal and the destination frames, and summary size K, we ask to find a summary $S =< \hat{F}, \hat{\delta} >$, consisting of an F-SummaryTree $\hat{F}$ and an δ-SummaryTree $\hat{\delta}$, such that (i) $\hat{F} \subseteq \Psi$, (ii) $\hat{\delta} \subseteq \Psi$, (iii) $|\hat{F}| + |\hat{\delta}| \leq K$, and (iv) the distance $d((F, \delta), (\hat{F}, \hat{\delta}))$ is minimized.

The distance function $d((F, \delta), (\hat{F}, \hat{\delta}))$ measures the quality of the summary. That is, it measures how well the selected cells summarize F-Tree and δ-Tree.

In our example a possible summary is: F-SummaryTree: {US_Northeast (235), NY (100), Queens (80), New_Haven (50)}, and δ-SummaryTree: {US_Northeast (-75), MA (-185), New_York (-100), Queens (-80)}.

4.1 Summary Loss Measure

To quantify the informativeness of a summary, we define a loss function that describes how difficult is to recover the original F-Tree and δ-Tree using the summary S. Intuitively, we want to reconstruct each cell value using the closest element on the summary, that is, we estimate the value of the cell using the closest super-cell (cell) in the summary. A super-cell distributes its score (count) to all its children, where each child is weighted based on its historical overall postings activity. For example, in the above F-SummaryTree, where the score for NY state is 100, if we assume that its three children (NY county, Bronx, Queens) have equal historic amount of postings, the estimation of score for each of them would be $100/3$.

To guarantee that all nodes are covered we assume that the root of Ψ is always included. To quantify the loss we use the following model. Given a cell r_i on the F-Tree F, and n is *any* entry in the F-SummaryTree $\hat{F}$ we define the cell loss $\L(r_j, n)$ as:

$$\L(r_i, n) = \begin{cases} 0 & \text{if } r_i = n, \text{ and} \\ |F(r_i) - \text{est}(r_i, n)| & \text{if } r_i \in R(n) \end{cases}$$

where $\text{est}(r_i, n)$ is the estimation for the score of r_i using the score given to n by the summary F-SummaryTree $\hat{F}$. As mentioned before, we only want to calculate the loss against the closest super-cell n that covers r_i. Let $\text{closest}(\hat{F}, r_j)$ return the closest cell such that r_i is in $R(\text{closest}(\hat{F}, r_j))$. Then the distance between the F-Tree and F-SummaryTree is defined as:

$$d(F, \hat{F}) = \sum_{r_i \in F} \L(r_i, \text{closest}(\hat{F}, r_i))$$

Similar definitions hold to calculate the distance between the δ-Tree and δ-SummaryTree.

Example 4.5 *Going back to the δ-Tree on Figure 4 and considering the differential component on our example summary, we can calculate the error as follows: NY and Queen County are covered without errors. Bronx's loss is $|5 - \text{est}(Bronx, NE)|$ units as it is covered by the Northeast (NE) region. New Haven (NH) loss is $|10 - \text{est}(NH, NE)|$ by the same argument.*

Finally, we extend can define the distance of the summary $S =< \hat{F}, \hat{\delta} >$ to the F-Tree F and δ-Tree δ as:

$$d((F, \delta), (\hat{F}, \hat{\delta})) = d(F, \hat{F}) + d(\delta, \hat{\delta}) \qquad (1)$$

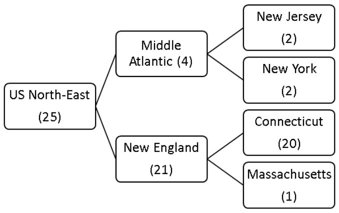


Figure 5: Example of non-monotonicity.

4.2 Summarization Algorithm

In this section we present an efficient algorithm to compute the optimal summary for a pair of frames, with respect to the loss function defined in Section 4.1. Our problem and solution have similarities to the *k-median problem.* Given a graph, the k-median problem looks for a set of k centers to open in order that minimize the sum of the distances for all the nodes in the graph to their closest center. This problem is **NP**-hard for general graphs but tractable for directed trees (Chrobak, Larmore, and Rytter 2001; Vigneron et al. 2000).

Note that our cost function is different to the one used for the k-median problem. The main problem is that the *loss* function presented in Section 4.1 is not a proper distance. In particular, the loss is non-monotonic as we go upwards on the tree, e.g., a node can sometimes be more accurately estimated (covered) by a farther node than by a closest selected node in the summary. Figure 5 shows an example of non-monotonicity. Consider Connecticut state. If we cover it using the New England cell, the loss would be $|20 - 21/2|$ $= 9.5$. If we use the root of the tree then the loss would be $|20 - 25/4| = 13.75$. In this case we would prefer the closest node. Now for the node Massachusetts the same analysis gives loss of 9.5 for the direct parent and 5.25 for the root. In this case we would prefer to use the root as the cover (and hence in the summary). In the k-median formulation we always prefer the closest node.

Our solution is based in the dynamic-programming bottom-up algorithm presented by Vigneron et al. (Vigneron et al. 2000). Let β be a node, and let anc be the closest ancestor of β (including β) that will be included in the summary. Then, $\text{cost}^j_{\text{anc}}(\beta)$ is the optimal selection of j nodes on the tree rooted at β assuming that anc is the closest selected ancestor. If $c_1, \ldots, c_m$ are the children of β, our dynamic-programming equation is:

$$\text{cost}^j_{\text{anc}}(\beta) = \min(\text{costInclude}(\beta), \text{costIgnore}(\beta))$$

where

$$\begin{aligned} \text{costInclude}(\beta) &= \min_{j_1+j_2\cdots+j_m=j-1} \sum_{j_k \in \{j_1 \cdots j_m\}} \text{cost}^{j_k}_{\beta}(c_k)\,, \\ \text{costIgnore}(\beta) &= \min_{j_1+j_2\cdots+j_m=j} \sum_{j_k \in \{j_1 \cdots j_m\}} \text{cost}^{j_k}_{\text{anc}}(c_k) \end{aligned}$$

Node β is added to the summary if the best way to select $j - 1$ elements on the sub-trees assuming β as a closer node (option: $\text{costInclude}(\beta)$) is better than selecting j elements on the sub-trees with anc as the closest node (option: $\text{costIgnore}(\beta)$). The base cases are set as follows: if we add

Table 2: Criterias to select hashtags

Strategy	Example
NBA Teams	lakers, heats,celtics,hawks,spurs
Musicians	gaga, bieber, nickelback, kellyclarkson
2012 Presidential Race	gingrich, ronpaul, romney, santorun, obama
Climate	snow, weather, climate, cold, storm
Cities	boston, nyc, seattle, philadelphia, atlanta
Trending	ettajames, concordia, costa concordia

a leaf node to the summary then the cost would be zero. If we do not select, we use the approximation error given by our loss function, i.e., $\text{cost}^0_{\text{anc}}(\beta) = |F(\beta) - \text{est}(\beta, \text{anc})|$. The complete solution is given by the $\text{cost}^j_{\text{root}}(root)$ entry

In spite of the non-monotonicity property, the above algorithm produces the optimal result for the following reasons: (i) selecting nodes in one sub-tree does not affect the selection of the sibling sub-trees — the reason is that our loss function only depends on the closest selected node; (ii) the algorithm can re-adapt the selected nodes as it keeps always the solutions for all ancestors. This avoids sub-optimal decisions, as β is only added to $\text{cost}^j_\beta(\text{anc})$ if we are sure we can not improve the solution using anc on a later step. We did not include the pseudocode due to lack of space and since the bottom-up dynamic programming tree navigation is relatively straightforward.

Distributing budget K: Given the algorithm to compute the best F-SummaryTree(or δ-SummaryTree) for any value of j, we consider the following strategies to split the budget K: the naive idea is assign the same number of elements to each tree ($k_1 = k_2 = K/2$). We call this strategy EQ-SPLIT. The second idea is to divide the budget in two parts assigning k_1 elements to the F-SummaryTreeand k_2 elements to the δ-SummaryTree, in such way that $k_1 + k_2 = K$ and the cost is optimal. Notice that the algorithm already calculates the possible costs so the combination is straightforward. We call this strategy SUM-k.

5 Evaluation

5.1 Data and Methodology

We use the Twitter stream API and follow 200 hashtags between Feb 1, 2012 and Feb 29, 2012. Half of the hashtags were selected because they were trending on the initial date. The rest were manually selected based on popular topics in the United States during the observation period. In particular, we selected hashtags about popular bands, sport teams, political figures, national and international issues and US cities. We collected 20.9 million tweets. Table 2 shows examples of the selected hashtags

The location of each tweet is determined using one of the two following heuristics. If a tweet is explicitly geo-tagged, we use this location. Else, we use the location string on the user profile. As Twitter does not require this location to be a valid value; we focus on those strings that match a strict regular pattern (CITY, STATE). We check that both the city and state appear in a locations dictionary. Given that user provided locations are noisy (Hecht et al. 2011), we expect this conservative heuristic to achieve reasonable precision but a low recall. In Section 6 we discuss other techniques that can be used to improve the initial location. Our final dataset has 850,000 geolocated tweets.

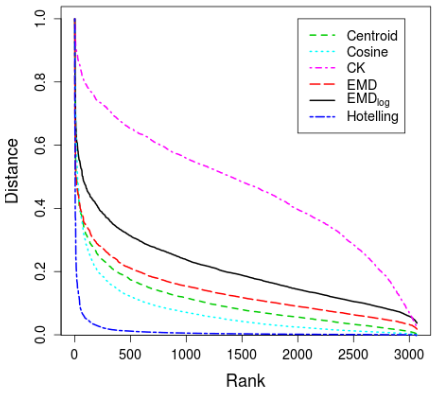


Figure 6: Distance vs. rank

For the geographical hierarchy we used the US Census hierarchy[1] defined as follows: county, greater city, state, division, region, country. Given the coordinates for a tweet we decide if it falls in some county boundaries.

Geometric distances between two points are measured using Euclidean distance. For areas, we use the distance between their centroids. All the frames used in the experiments are 24 hours long, i.e., we compare the activity of two different days. In this way, we avoid the problems of temporal changes on the activity between different times of the day. A frame is valid if it has at least 300 tweets. Our frame dataset consists of 3048 pairs, i.e, two valid frames on consecutive days for the same hashtag.

5.2 Movement Measure Analysis

We compare the measures proposed in Section 3. The first experiment measures the correlation among the proposed measures. The second experiment is a qualitative evaluation of the measures. Finally, we show examples of movements.

Quantitative Analysis

Distribution of Movement Scores: This experiment studies how the movement scores change for each movement measure. We consider all the 3 048 pairs of frames and rank then the pairs by score. Figure 6 shows how the score changes against the rank .The x-axis is the rank, that is, the leftmost point corresponds to the frames pair with maximum movement. All measures are normalized between $[0, 1]$, in such way that higher values imply higher movement.

Excluding the CK measure, all the proposed measures drop very quickly after the first hundred elements and become indistinguishable on the lower ranks. Another observation is related with the speed of the drop itself. We see that Hotelling is the most steep, which is undesirable. On the other hand, the CK and EMD drop is slower, favoring element difference.

An interesting observation is that the line of CK measure can be divided into 3 parts: interesting movements (steep left

[1] http://www.census.gov/geo/www/tiger/

Table 3: Kendall-τ distance

	$\mathrm{EMD}_{\log}$	Cosine	Centroid	Hotelling	CK
EMD	0.77	0.64	0.65	0.21	0.25
$\mathrm{EMD}_{\log}$		0.77	0.52	0.10	0.30
Cosine			0.42	0.05	0.43
Centroid				0.36	0.16
Hotelling					0.08

part), average movement (middle slowly decreasing part), and low movement (steep right part). In contrast, the others display only two drop rates, i.e., only one elbow. We can set a threshold to distinguish when an object is moving or not. This would be the value at the elbow as frames after that point are barely different.

Similarity among movement measures: We compare the different measures against each other, to understand how similar they are. Again we rank 3 048 pairs of frame and measure the difference between rankings. We use the Kendall-τ distance, which describes the probability that the pair-wise order of elements x, y agree in both rankings.

As we see in Table 3 the two measures of the EMD family are highly correlated. Interestingly, the correlation values are also very high with the cosine and centroid measure. Moreover, we can define a classification of the measures. As the Cosine and CK measure do not consider the distance, they are highly correlated with $\mathrm{EMD}_{\log}$. On the other hand, the standard EMD is more correlated with the centroid measure. Hotelling is uncorrelated with all the other measures.

Based on these correlations we partition the proposed measures into three groups:

- Distance-sensitive measures: EMD, $\mathrm{EMD}_{\log}$ and Centroid depend on the distance of the movement
- Non-Distance Sensitive Measures: Cosine, CK are not related with the distance. Interestingly enough the $\mathrm{EMD}_{\log}$ measure is correlated with both.
- Hotelling: Hotelling does not behave as the previous measures.

Qualitative Analysis: User evaluation.

To evaluate the proposed measures we perform a MechanicalTurk user study. The goal is to test which techniques are more consistent with the human perception of movement.

To visualize the movement we create two-frame animations of movements. To create each image frame we use the same heat-map methodology described in Section 3 (darker color means more posts). Frames are normalized to have the same number of tweets. Examples of the two frames of an animation are shown in Figure 1.

To decide which measure is better we focus on those movements for which the pair-wise rankings disagree, i.e., they appear in significantly different positions in two rankings measures. We present these conflicting pairs to the user, so she can select which one she thinks has a clearer movement. We expect users to prefer those animations that contain a coherent and clear visual pattern, and reject those that are noisy.

To select the pairs of movements we use the following methodology: We first pick an animation in the ranking of the first measure and a second animation in the ranking of the second measure. We check that the difference between the positions on both rankings is at least 100 positions. This ensures strong disagreement (controversy) between the measures on the same pair.

As comparing all pairs of measures is too expensive, we consider the grouping presented before. We compare a distance measure representative, a non-distance representative and the Hotelling measure. In particular we compare the $\mathrm{EMD}_{\log}$, centroid, CK and Hotelling. For each pair of measures we select 30 random "controversial" pairs.

We recruited 50 Mechanical Turk users. For each user, we present a survey of 20 animation pairs. [2]. Each animation is presented to five users and we select the best measure for that pair by a simple majority vote. Figure 7 shows the results. We see that the Hotelling measure is the clear loser. We also see that $\mathrm{EMD}_{\log}$ and CK measures are the ones that are preferred by the users. Both measures are clearly better than the simple centroid measure in most of the cases. As we see the differences are significative enough to be trusted.

Comparing centroid and $\mathrm{EMD}_{\log}$ we observe that the later seems to be better only by a small margin. The reason is that these two measures are highly correlated as they are based on distance. The clear winner is the CK measure, which captures the idea of non-overlapping changes, which seem to be more natural for users. Nevertheless, for the CK measurem movements from one city to another are as important as movements between coasts. However, we think that distance can be meaningful for some applications, so we do not discard the distance based measures.

Visual Inspection of the Movement.

To better understand the performance of these methods, Table 4 presents a sample of 10 movements detected by the $\mathrm{EMD}_{\log}$ and CK measures, that is, these movements received a high score for these measures. For each movement we describe the hashtag and location shift between consecutive days, and the intuitive description of the change, which we generate manually by studying the relevant news.

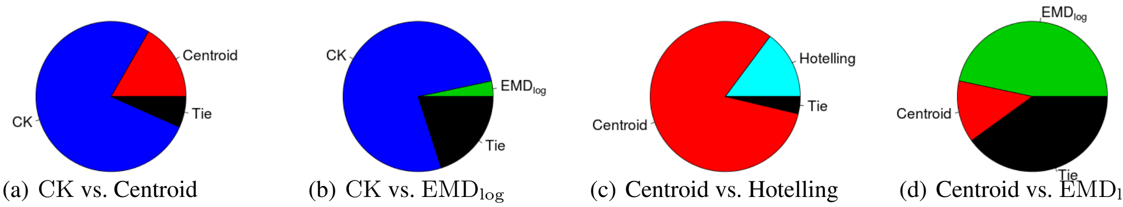
We see that the detected movements correspond to interesting movements in the real world. For example, climate tags like **`#Snow`** and **`#Storm`** are related with corresponding events. Similarly, sport teams (**`#Celtics`**) and music bands (**`#Ladyantebellum`**) move as they travel around the country. We also see that if the same hashtag is used for different topics, the movement detection can be used to detect the noise, e.g., **`#Concordia`** refers usually to a college team but also briefly to Costa Concordia.

5.3 Summarization Evaluation

In this section we evaluate the summarization strategy proposed in Section 4. We use the US administrative hierarchy. We weight each county using the total number of posts. We compare the following strategies:

- BEST-COUNTY: baseline that selects the K counties (leafs) that give the smallest loss.

[2] Some of the presented pairs were used as validation to eliminate users who make random selections



(a) CK vs. Centroid (b) CK vs. EMD$_{\log}$ (c) Centroid vs. Hotelling (d) Centroid vs. EMD$_{\log}$

Figure 7: User survey results: preferred method in pair-wise comparisons.

Table 4: Movement examples.

HashTag	Date	Location	Description
CNNDEBATE	20120224/25	NY,SF to Dallas,TX	Country wide burst for the GOP Debate. Next day a single user spams with the tag.
SNOW	20120207/08	SF, LA, Colorado to New England	Snow Storms in Colorado, and discussion of sky conditions in California. Next day, there are storms in NYC and Colorado.
SNOW	20120220/21	NC/Virginia to Midwest Cities	Snows in Virginia and North Carolina. Next day, snow in Minneapolis, Chicago
LADYANTEBELLUM	20120129/30	Des Moines, IA to Bloomington, IL	Lady AnteBellum plays in Des Moines and Bloomington
STORM	20120225/26	New York, Illinois, Seattle to US	Storms in Upstate New York. Next day there are storms all around the country.
LADYANTEBELLUM	20120215/16	SLC (Utah), Denver (CO) to Denver (CO)	Lady Antebellum plays at 14 in SLC, and people comment at 15. A second show is scheduled at 16 in Denver.
CELTICS	20120208/09	Boston to Boston, LA	Celtics rest at 8th and play next day against the LA Lakers (88-87).
BACHMAN	20120220/21	US to Minneapolis	Michelle Bachman decides to go for the 6th seat in Minneapolis
SPURS	20120206	San Antonio, NYC, LA, SF	San Antonio Spurs in home city. Next day, Liverpool plays against Tottenham Spurs
CONCORDIA	20120220	St Paul to Big Cities	Concordia St at St Paul college sport tweets. Next day Discover special on Concordia Accident

- BEST-STATE: baseline that selects the K states that give the smallest loss.
- EQ-SPLIT: described in end of Section 4.2.
- SUM-k: described in end of Section 4.2.

We compare these strategies in terms of reconstruction and coverage. The reconstruction error is defined by Equation 1. The coverage is the percentage of activity that is covered by super-cells (except the root).We calculate summaries of size $K = 4$, 8, 16 and 32 entries, for 300 hashtag-days pairs (movement of a hastag for two consecutive days). Figure 8 shows the error for each strategy. The best-state strategy is the worst as it does not capture activity in specific areas. On the other hand, our proposed strategies are on average 35% better than the best-state strategy and 15-25% than the best-county strategy.

Figure 9 shows the coverage of each strategy at a specific level of the hierarchy, e.g., the state bars show the coverage if we only consider those nodes in the summary that are more specific than a state. Ideally we can cover all the entries picking elements on the higher levels. But this would make the estimates very coarse. For example, the best-state strategy covers at least 80% of the tweets, but has high reconstruction error. In contrast the best-county strategy has the lowest coverage (70%). Our proposed strategies have 75% coverage. This difference is explained as our summaries include more general regions only when necessary.

Finally, we note a small difference between sum-k and eq-split. In our experiments, sum-k allocates on average about 70% of the budget to the base frame and only 30% to the delta frame, but this does not seem to make a difference to the reconstruction error.

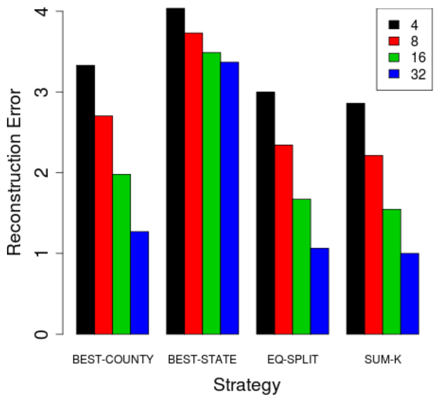


Figure 8: Reconstruction error

6 Related Work

Location in Microblogging: Mathioudakis et al. (2010) present an algorithm to identify regions that are active or bursting in a particular period of time. We focus on movements or changes on different days instead of the change of activity for a fixed region. Mehler et al. (2006) present a spatial analysis of news sources, visualizing with heat maps mentions of named entities in US newspapers. In contrast, our work deals with the visualization of spatial movement and considers real-time sources.

Backstrom et al. (2008) present a parametric model to detect the center of activity in query logs. They show that change can be related with real event movements. (Sakaki, Okazaki, and Matsuo 2010) use a different approach to show that Twitter users can be considered as sensors, that react to spatial static events or moving patterns. Our work focuses on measuring and summarizing this movement. (Eisenstein

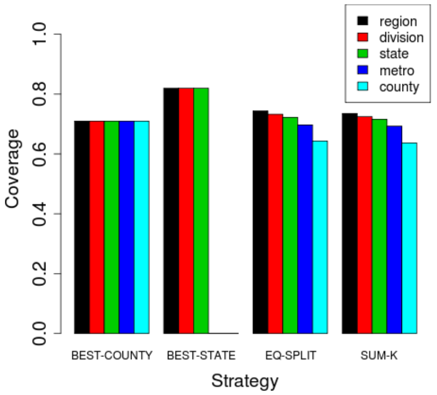


Figure 9: Coverage error

et al. 2010) present a model to jointly detect topics and location of tweets, recognizing the regional language used for each topic. (Cheng, Caverlee, and Lee 2010), (Ikawa, Enoki, and Tatsubori 2012) present methods to estimate the location of posts. We can leverage these works for topic and location estimation, which are problems orthogonal to our work.

EMD: EMD has been used in Information Retrieval for image similarity (Rubner, Tomasi, and Guibas 2000). In theory EMD can be solved using the simplex algorithm, but in practice it has been shown to be super-cubic for this kind of systems (Ling and Okada 2007). Faster approximations and approximations have been proposed in implementation and approximations (Ling and Okada 2007; Pele and Werman 2009).

7 Conclusions

We studied the problems of measuring and summarizing geo-spatial movement of microblog hashtags. We proposed and compared three classes of approaches to quantify movement and showed that our measures can be used to decide when geo-spatial changes are meaningful. We proposed an efficient movement summarization algorithm that creates a summary composed of two parts: an initial description of the hashtag map before the movement and a delta-frame that describes the change. We showed that our multi-granular summary improves over baselines in both coverage and reconstruction error.

Acknowledgements.

This work was supported in part by National Science Foundation grants OISE-0730065, IIS-1216032 and IIS-1216007.

References

Backstrom, L.; Kleinberg, J.; Kumar, R.; and Novak, J. 2008. Spatial variation in search engine queries. In *Proc. WWW'08*, 357–366. New York, NY, USA: ACM.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Campana, B. J. L., and Keogh, E. J. 2010. A compression based distance measure for texture. *Stat. Anal. Data Min.* 3(6):381–398.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*, 759–768.

Chrobak, M.; Larmore, L.; and Rytter, W. 2001. The k-median problem for directed trees. In *Mathematical Foundations of Computer Science 2001*, volume 2136. 260–271.

Conover MD, Davis C, F. E. M. K. M. F. e. a. 2013. The geospatial characteristics of a social movement communication network. *PLoS ONE* 8(3).

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *Proc. of EMNLP '10*, 1277–1287.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

González, M.; Hidalgo, C.; and Barabási, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779.

Hazas, M.; Scott, J.; and Krumm, J. 2004. Location-aware computing comes of age. *Computer* 37(2):95–97.

Ikawa, Y.; Enoki, M.; and Tatsubori, M. 2012. Location inference using microblog messages. In *Proc. WWW '12*, 687–690.

Lappas, T.; Vieira, M. R.; Gunopulos, D.; and Tsotras, V. J. 2012. On the spatiotemporal burstiness of terms. *PVLDB* 5(9):836–847.

Li, M.; Chen, X.; Li, X.; Ma, B.; and Vitányi, P. 2003. The similarity metric. In *SODA*, 863–872.

Ling, H., and Okada, K. 2007. An efficient earth mover distance algorithm for robust histogram comparison. *PAMI* 29:853.

Mathioudakis, M.; Bansal, N.; and Koudas, N. 2010. Identifying, attributing and describing spatial bursts. *Proc. VLDB Endow.* 3:1091–1102.

Mehler, A.; Bao, Y.; Li, X.; Wang, Y.; and Skiena, S. 2006. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics* 12:765–772.

Nasrabadi N.M. Lin, S. F. Y. 1989. Quad tree structures for image compression applications. *Proc. ICASSP '89*.

Pele, O., and Werman, M. 2009. Fast and robust earth mover's distances. In *ICCV '09*.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40:99–121.

Ruiz, E. J.; Hristidis, V.; Castillo, C.; Gionis, A.; and Jaimes, A. 2012. Correlating financial time series with micro-blogging activity. In *WSDM*, 513–522.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 851–860.

Vieira, M. R.; Bakalov, P.; and Tsotras, V. J. 2010. Querying trajectories using flexible patterns. In *EDBT*, 406–417.

Vigneron, A.; Gao, L.; Golin, M. J.; Italiano, G. F.; and Li, B. 2000. An algorithm for finding a k-median in a directed tree.

Zhang, H.; Korayem, M.; You, E.; and Crandall, D. J. 2012. Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities. In *WSDM*, 33–42.