Public Checkins versus Private Queries: Measuring and Evaluating Spatial Preference

James Caverlee*, Zhiyuan Cheng*, Wai Gen Yee†, Roger Liew†, Yuan Liang*
*Texas A&M University, †Orbitz
*{caverlee, zcheng, yliang}@cse.tamu.edu, †{waigen.yee, roger.liew}@orbitz.com

ABSTRACT

Understanding the spatial preference of mobile and web users is of great significance to creating and improving location-based recommendation systems, travel planners, search engines, and other emerging mobile applications. However, traditional sources of spatial preference – which reflect the patterns of geo-spatial interest of large numbers of users – have typically been expensive to collect, proprietary, and unavailable for widespread use. In this paper, we investigate the viability of new publicly-available geospatial information to capture spatial preference. Concretely, we compare a set of 35 million publicly shared check-ins voluntarily generated by users of a popular location sharing service with a set of over 400 million private query logs recorded by a commercial hotel search engine. Although generated by users with fundamentally different intentions, we find common conclusions may be drawn from both data sources – (i) that the relative geo-spatial "footprint" of different locations is surprisingly consistent across both; (ii) that methods to identify significant locations results in similar conclusions; and (iii) that similar performance may be achieved for automatically identifying groups of related locations. These results indicate the viability of publicly shared location information to complement (and replace, in some cases), privately held location information.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; J.4 [**Computer Application**]: Social and Behavioral Sciences

General Terms

Algorithms, Experimentation

Keywords

Location-based services, checkin, queries, spatial data mining

1. INTRODUCTION

Social scientists and geographers have long been interested in modeling the linkages and flows between locations for better understanding a variety of geo-spatial issues including: why and how migration flows among countries, regions, and cities; to model commerce flows and explain trade relations among trading partners; to design more efficient roadways and traffic forecasting; to develop epidemiological models of disease spread; and so forth. This *spatial interaction* is a cornerstone of geographic theory, "encompassing any movement over space that results from a human process" [9]. Traditional methods for modeling these flows and the *spatial preference* of users in one location for another location have typically relied on expensive and hard-to-maintain data sources, like the 10-year US Census, which collects massive statistics about the connections between people and between cities in the United States.

As a point of excitement, the rise of the web over the past ~20 years has seen a commensurate rise in the low-cost collection of implicit linkages and flows among users and locations. For example, millions of people share their location information passively while using on-line services like video streaming services (e.g., Amazon Instant Video, and Netflix), search engines (e.g., Google, Bing), e-commerce sites (e.g., eBay, and Amazon), and travel planning sites (e.g., Orbitz, Expedia, and Priceline). By tracking IP addresses, plaintext queries, and other location identifiers, these proprietary services have been harvesting huge databases of spatial interaction. For example, by aggregating user search and purchase decisions, Amazon can identify the interest level of users in one location for another location (e.g., more customers in California are buying Texas guidebooks, which may be an early indicator of future migration). However, the excitement over these sources of spatial interaction must be tempered by the proprietary nature of the data.

Fortunately, the past few years have seen the widespread *voluntary* sharing of location information by users of location-sharing services. As GPS-enabled devices have become ubiquitous, users of services like Twitter, Foursquare, and Gowalla have begun actively sharing fine-grained spatial information about their life, interests, and footprints in real-time. This voluntary sharing provides unprecedented opportunities to study people in different regions, and the connections between people and places. In comparison with expensive, proprietary, and often times unavailable resources, this publicly-shared data offers the promise of new methods appealing not only to geographers and social scientists, but to computational researchers and practitioners seeking to create and improve location-based recommendation systems, travel planners, search engines, and other emerging mobile applications.

Hence, in this paper, we investigate the viability of new publicly-available geospatial information to capture spatial preference. Concretely, we explore the spatial preference of users from two large-scale datasets: a set of private query logs for hotels automatically recorded by a commercial on-line hotel search engine (Orbitz), and a set of publicly available check-ins voluntarily generated by users from a typical location sharing service (Gowalla). The check-in

ACM SIGSPATIAL LBSN'12 November 6, 2012. Redondo Beach, CA, USA
Copyright 2012 ACM ISBN 978-1-4503-1698-9/12/11 ...$15.00.

data includes over 35 million check-ins from 1.2 million users from Gowalla. The hotel query log data includes all the queries and bookings for hotels from the hotel search engine in 2011, which in total includes over 400 million records from over 20 million unique IPs. We explore in this paper the commonalities and the differences between these two sources of spatial preference – generated by different user bases with fundamentally different intentions.

Concretely, this paper makes three contributions:

- First, we model the spatial preference of users across both datasets and measure the relative geo-spatial "footprint" of different locations via three localness metrics: the mean contribution distance, the radius of gyration, and the city locality. We find that though the absolute values of these metrics differ across datasets, the relative values are surprisingly consistent.
- Second, we develop a PageRank-like method for identifying spatially significant locations based on the spatial preference of users. Through a random walk over the spatial preference graph linking locations, we find that both datasets reveal similar significant locations.
- Third, we investigate the potential of mining related clusters of locations from both datasets based on the spatial preferences of users. In a comparison against a ground truth of 800 hand-curated lists of related cities, we find similar performance across both public and private datasets.

These results indicate the viability of publicly shared location information via checkins to complement (and replace, in some cases), privately held location information such as that in proprietary query logs. The potential of publicly shared location information serving as a substitute for privately held information could provide new avenues of research for social scientists, geographers, as well as computer scientists interested in the geo-spatial flows of ideas, memes, and geo-targeted applications.

2. RELATED WORK

Researchers have been investigating the spatial properties of large-scale data for many years. In the context of query logs, there have been several efforts typically targeted at the spatial properties revealed through text-based queries to large search engines. For example, Backstrom et al. [2] introduced a model of spatial variation for analyzing the geographic distribution of queries using Yahoo's query logs. The authors proposes a generative probabilistic model in which each query has a geographic focus on a map (based on an analysis of the IP-address-derived locations of users issuing the query). Gan et al. [7] conduct an analysis of 36 million queries from AOL, and identified typical properties for queries with a geographic intention. In addition, they built a classifier that can accurately classify queries into geographic and non-geographic queries.

With the rise of online social networks, there has been a similar rise in analyzing the spatial patterns revealed. For example, Facebook researchers [3] observed that Facebook users have more local friends than distant friends, and that they can predict a Facebook user's location with high accuracy given the location for the users' friends. McGee et al. [12] investigate the relationship between the strength of the social tie between a pair of friends and the distance between the pair with a set of 6 million geo-coded Twitter users and their social relations. They observed that users with stronger tie strength (reciprocal friendship) are more likely to live near each other than users with weak ties. Hecht et al. [10] study the localness of user generated content in Flicker and Wikipedia, and they observe that the content generated by Flickr users is more local comparing to the content generated by Wikipedia editors. A host of related work has also focused on mining interesting trajectories [19], modeling periodic behaviors and mobility patterns [5, 8, 4], and studying the correlation between people's social relations and their mobility patterns [6]. Others have focused on location recommendation at the point of interest (POI) level based on queries and bookings for hotels [13], and check-ins in location sharing services [17, 18, 16]. In the granularity of city-level, researchers have studied the interaction between cities via on-line social relations [11].

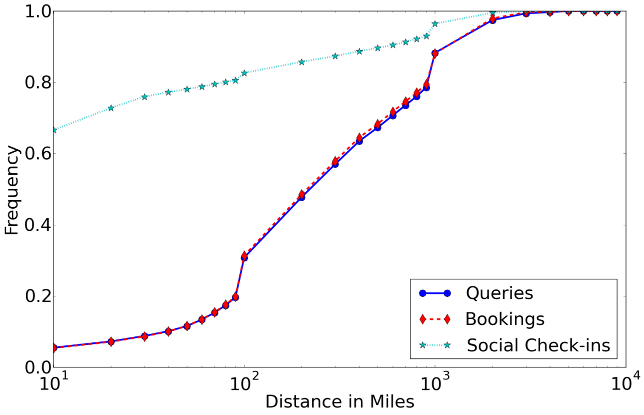


Figure 1: Distance versus Frequency: Check-ins tend to be more local; 80% of all check-ins are within 100 miles of a user's home location. In contrast, query (and booking) locations are more distant; only 25% are within 100 miles of a user's home location.

3. DATA

As the basis of this investigation, we consider two large-scale datasets: a set of private query logs and a set of publicly available check-ins.

3.1 Private Spatial Resource: Query Logs

The hotel query log data includes a large set of both queries and bookings for hotels randomly sampled from a commercial on-line hotel search engine – Orbitz. The dataset includes over 400 million records, from over 20 million unique IPs all over the world. Each query (or booking) includes an IP address which can be translated to a city-level location where the query (or booking) is issued. We call this the origin location. Each query (or booking) also contains another city-level location indicating the destination (i.e., the city where the queried hotel is located).

To focus on legitimate users of the Orbitz search engine, we filter out IP addresses accounting for an anomalous number of searches (greater than 2,000 queries each). For example, several thousand IPs generate from thousands to millions of queries each; most likely, these are search engine crawlers or bots from other travel search engines). Additionally, we focus on queries (and bookings) originating from the Continental United States. Considering each unique IP as a unique user, we consider the corresponding city-level location for the IP as the home location for the user, resulting in **69 million queries and 1.1 million bookings**.

3.2 Public Spatial Resource: Check-ins

The check-in dataset[1] includes over 35 million check-ins from about 1.2 million users from Gowalla, a popular location-sharing service. Each of the check-ins includes a fine granular point of interest (POI) location (i.e., where the check-in happened), a timestamp (i.e., when the check-in happened), and a piece of short text (i.e., what the check-in is about). Each check-in's POI location links to a particular city, which allows us to group the check-ins

[1] Please refer to http://infolab.tamu.edu/data for the dataset.

into city-level locations. For each user, we simply consider the city which has the most check-ins from the user as the home location.

Similar to the query log data, the check-in data also reveals each user's interest in other "destinations", in this case by considering check-in locations outside of the user's home location. For example, a user from Los Angeles who checks-in in New York City indicates that user's interest in New York. As in the case of the query log data, we focus only on locations within the Continental United States and we filter out users with fewer than 20 check-ins each. The filtering leaves a set of almost 70,000 users and over **15 million check-ins** from the users.

3.3 Private versus Public

These two resources – one private and one public – are naturally quite different. Users of these two services vary in their demographics since location sharing service users tend to be young with access to a mobile device, while hotel search engine users are more often representative of the general public with access to a desktop computer. And of course, users of these two services have fundamentally different goals. Hotel queries reflect a user's future intent; check-ins reveal a user's current physical movement. Hotel query logs are more likely to reveal long-distance travel intentions, whereas check-ins are typically a more local phenomenon reflecting a user's interest in local restaurants, bars, and stores [5]. Users of location sharing services are also intentionally sharing their location information, whereas users of search engines are not consciously sharing their location with others (though these search engines may log and analyze the user's queries, IP address, and other location-revealing artifacts). With these many differences in mind, we next turn to an investigation of the spatial preference embedded in these two sources and whether we can find any commonalities between them. Finding such commonalities could demonstrate the potential of publicly shared location information serving as a substitute for privately held information.

4. EXPLORING SPATIAL PREFERENCE

We begin our investigation by exploring the spatial preference revealed through both datasets. We model the spatial preference and measure the relative geo-spatial "footprint" of different locations via three localness metrics: mean contribution distance, radius of gyration, and city locality.

4.1 Preliminaries

Each query (or booking) in the private dataset and each check-in in the public dataset reveals a bidirectional relationship between an origin location and a destination location. In the case of queries (or bookings) the origin is the city-level location of the user issuing the query; the destination is the city-level location of the hotel. In the case of the check-ins, the origin is the user's home location (which we define as the city with the most check-ins by the user); the destination is the city-level location of the current check-in.

To start with, we are interested in investigating the basic properties of these origin-destination relationships. For each set of queries, bookings, and check-ins, we bucket all the distances between origins and destinations into groups. Figure 1 plots the cumulative frequency of the pairs of origin and destination bucketed into groups of distance. The patterns of the bookings and the queries are almost identical to each other, with over 5% of the queries (bookings) for hotels within 10 miles, and about 30% within 100 miles. On the other hand, the check-ins are much more local comparing to the hotel queries (bookings). Over 65% of the check-ins are within 10 miles to the users' home locations, and over 80% are within 100 miles. This difference is our first sign that these two resources reflect fundamentally different usages: that people use hotel search engines to look for hotels to stay during their business trips or vacations, and people use location sharing services to share the real-time status of their daily activities.

Table 1: Average Value for Cities' Localness Metrics

Localness Metric	MCD (miles)	R_g (miles)	CL
Queries	869.346	549.904	0.560
Bookings	809.456	522.644	0.569
Check-ins	380.121	134.477	0.614

4.2 Spatial Preference

Given pairs of origin location and destination location extracted from queries (bookings) and check-ins, we quantify the *spatial preferences* for each of the cities with a spatial preference probabilistic distribution. Spatial preference is intended to reflect the aggregate interest level of users in an origin location for a particular destination location.

Spatial Preference: Let l_i be an origin location and let l_j be a destination location. Let $S(l_i)$ be a set of all pairs of origin-destination records in the dataset that originate from location l_i, and let $S(l_i, l_j)$ be a set which includes all pairs of origin-destination records that originate from location l_i with a destination in l_j. Then the spatial preference for location l_i toward location l_j is:

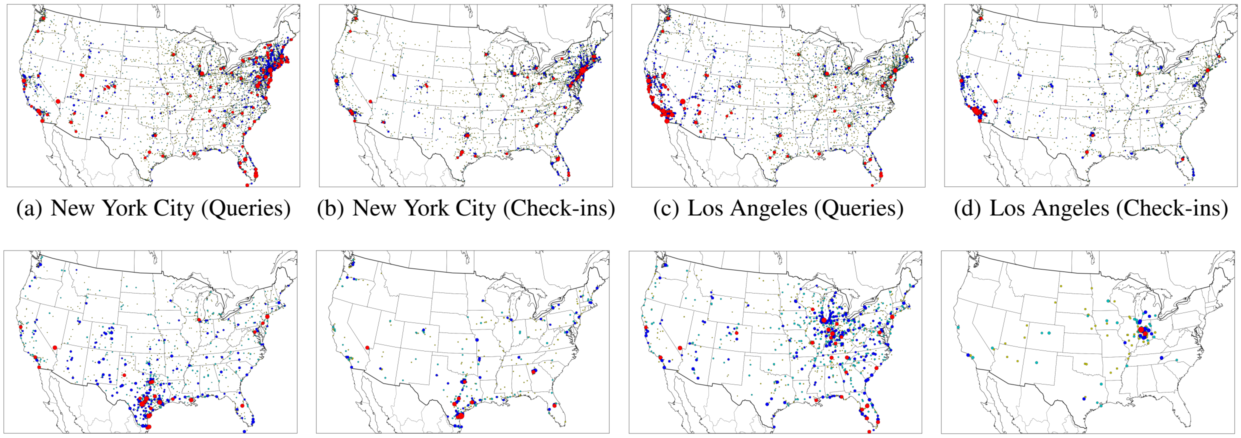
$$p(l_i, l_j) = \frac{|S(l_i, l_j)|}{|S(l_i)|}$$

Example: For example, suppose we have 10 total records (either from the query data or the check-in data) with an origin location of A. Of these, there are three occurrences of <A, A>, two occurrences of <A, B>, and five occurrences of <A, C>. Then, the spatial preferences for location A toward locations A, B, and C are: $p(A,A) = \frac{3}{10} = 0.3$, $p(A,B) = \frac{2}{10} = 0.2$, and $p(A,C) = \frac{5}{10} = 0.5$. Hence, users in location A have the strongest preference for location C, and the weakest preference for location B.

Given the definition of spatial preference, we map the spatial preference originating from four cities across both the private query data and the public check-in data. Figure 2 highlights the spatial preference of New York City, Los Angeles, Corpus Christi (Texas), and West Lafayette (Indiana). In each of the figures, the color and size of the dots indicate the intensity of the spatial preference from the origin to the destination: red indicates the top 2% most preferred cities; blue indicates the top 2% to 20%; cyan indicates the top 20% to 50%; and yellow indicates the bottom 50%.

As we observe in the figures, the private query data is much denser compared to the check-in data. This is partially an artifact of the data collection limits we faced but is also a reflection of the relative density of these two sources – query logs are inherently a much larger potential collection than are check-ins. Even with this difference in density, we note the relative similarity of the spatial preferences measured across source. People from New York are most interested in the northeast corridor; people from Los Angeles are most interested in the west coast; similar observations can be made for the much smaller locations of Corpus Christi and West Lafayette.

Additionally, we observe that queries balance their locality with many distant locations. For example, Figure 2(a) shows that New Yorkers have many queries for hotels in the New England area, but they are also interested to travel to the Florida and to the west coast. Similarly, Figure 2(c) also shows a a balance between local queries and for more distant ones. In comparison, the check-in data – though of a national scale for both New York and Los Angeles – is much more local (further confirming the relative localness



(a) New York City (Queries) (b) New York City (Check-ins) (c) Los Angeles (Queries) (d) Los Angeles (Check-ins)

(e) Corpus Christi (Queries) (f) Corpus Christi (Check-ins) (g) West Lafayette (Queries) (h) West Lafayette (Check-ins)

Figure 2: (Color) Spatial Preference for Example Cities. Figures in the left-column are derived from private query logs. Figures in the right-column are derived from public check-ins. The color and size of the dots indicate the intensity of the spatial preference from the origin to the destination: top 2% (red); 2-20% (blue); 20-50% (cyan); and the bottom 50% (yellow).

of check-ins versus queries in Figure 1). Queries for hotels are relatively more local for the two smaller cities, as we can see in Figure 2(e) and Figure 2(g). In comparison, the check-in spatial preferences are much sparser and more focused around the origin location.

4.3 Comparing Localness

Given the spatial preference probabilistic distribution for a specific location, we can describe each location by measuring its *localness*. The goal of such a localness measure is to encode the entire distribution of spatial preferences into a single summary metric. By evaluating each location, we can directly compare the localness of locations as described by private query logs and by public check-ins. Toward this goal, we adopt three complementary measures of localness:

Mean Contribution Distance (*MCD*)

Proposed by Hecht et al. [10], the *MCD* measures the weighted average of the distances between an origin location and multiple target locations:

$$MCD(l_i) = \Sigma_{l_j \in S} \left(\frac{d(l_i, l_j) * |S(l_i, l_j)|}{|S(l_i)|} \right)$$

where S includes all locations of interest and $d(l_i, l_j)$ denotes the distance between the origin location l_i and a target location l_j. A small value indicates strong localness for a city; most users in the origin location either query for or check-in to nearby locations. A large value indicates more global interest; users either query for or check-in to distant locations.

Radius of Gyration (r_g)

Adopted for location analysis by Gonzalez et al. [8], the r_g measures the standard deviation of distances between an origin location and target locations:

$$r_g(l_i) = \sqrt{\frac{1}{|S(l_i)|} \sum_{l_j \in S} (d(l_i, l_j))^2 * |S(l_i, l_j)|}$$

In essence, the radius of gyration measures both how frequently and how far people from the origin travel. A low r_g typically indicates a location whose residents travel mainly locally, while a high radius of gyration indicates a location with many long-distance travelers.

City Locality (*CL*)

The third measure of "localness" is city locality, proposed by Scellato et al. [14]. The city locality for city (l_i) is formally defined as:

$$CL(l_i) = \frac{1}{|S(l_i)|} * \sum_{l_j \in S(l_i)} |S(l_i, l_j)| * e^{-d(l_i, l_j)/\beta}$$

where β is a scaling factor used to normalize the values of localities so that city localities can be compared using different data and geographic sizes. The city locality is always normalized between 0 and 1. A city with high localness has a higher value of city locality. In practice, the scaling factor β is picked as the mean distance between all the pairs of spatial preference between different cities.

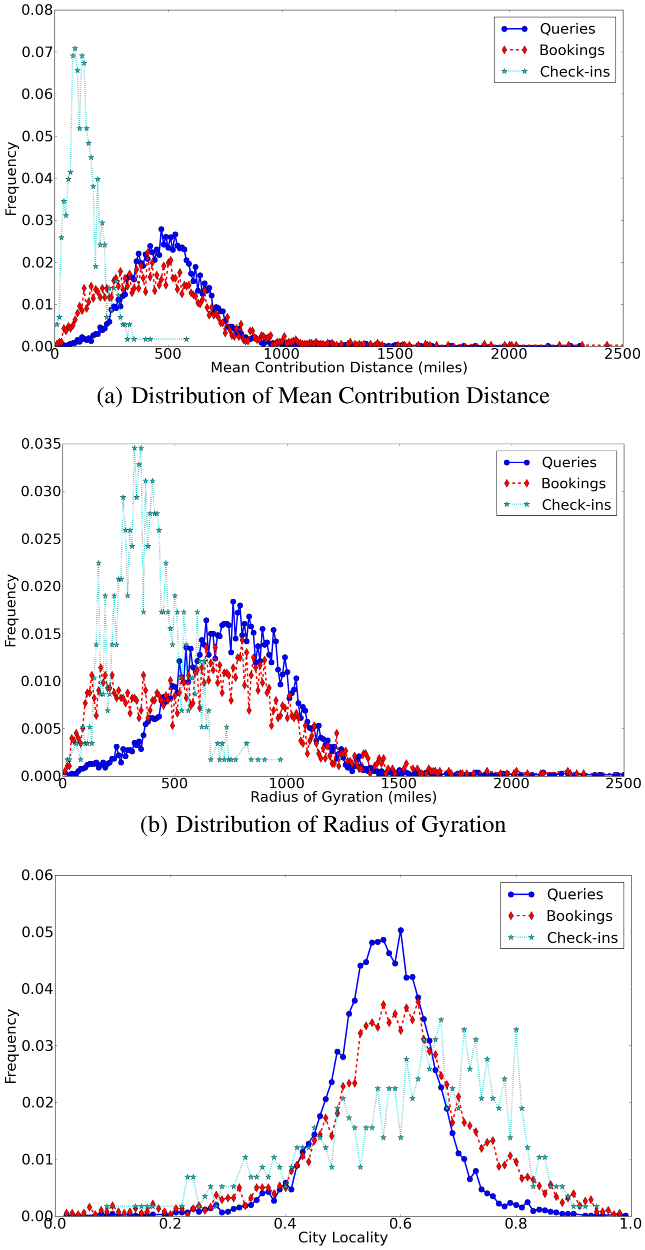
Provided the three localness metrics above, we compare the localness between different cities via their localness metrics. To calculate the localness metrics for each of the cities, we firstly filter out cities without dense data. Specifically, for queries (or bookings), cities with fewer than 1000 queries are filtered out. Similarly, for check-ins, cities with fewer than 1000 check-ins are filtered out. Based on the remaining cities, we calculate each of the three localness measures across queries, bookings, and check-ins.

Table 1 shows the average values of the three localness metrics for cities in the three datasets. We see that the private queries (and bookings) naturally reveal a larger scope of interest as compared to public check-ins. The MCD is around 400 to 500 miles greater; the radius of gyration is around 400 miles greater, and the city locality measure is lower (indicating less localness in comparison). Intuitively, it seems reasonable that check-ins are much more local since they are more constrained by physical mobility (e.g., I have to travel to the location, then reveal my location). As a side note, we see that queries are even less local than bookings, suggesting the exploratory possibility of querying, versus the reality of actually booking a hotel (e.g., it's fun to consider far-flung trips, but in actuality we tend to book more reasonable destinations).

Further confirming this intuition, we show in Figure 3, the complete distribution for each of the three localness measures across the private queries (and bookings) versus the public check-ins. We see that the distributions are approximately Gaussian with the check-in distribution resulting in smaller mean contribution distance and

Table 2: Values of Localness Metrics for Example Cities

Localness Metric	*MCD* (miles)			R_g (miles)			CL		
	Queries	Bookings	Check-ins	Queries	Bookings	Check-ins	Queries	Bookings	Check-ins
New York City	812.384	932.113	310.563	1278.979	1360.094	747.878	0.418	0.389	0.334
Los Angeles	627.814	619.859	174.568	1056.731	1017.116	541.458	0.551	0.552	0.637
Corpus Christi, TX	435.364	356.841	172.819	693.989	565.083	432.333	0.581	0.618	0.231
West Lafayette, IN	599.479	543.760	121.559	887.397	818.425	282.017	0.510	0.539	0.476



(a) Distribution of Mean Contribution Distance

(b) Distribution of Radius of Gyration

(c) Distribution of City Locality

Figure 3: Distribution of Localness Metrics

smaller radius of gyration, relative to the others. The city locality for check-ins is also skewed more rightward, again conveying the more localness of the check-in data. Connected to the earlier side note, we can see that the bookings are more local than queries based on their distributions.

Finally, we can revisit our four example cities – New York City, Los Angeles, Corpus Christi, and West Lafayette – in terms of the three localness metrics. As shown in Table 2, comparing to an average city, people from New York City really travel to a lot of distant cities even farther than the places they searched for. For Los Angeles, the bookings are only slightly more local compared to the queries, while the differences between bookings and queries for Corpus Christi and West Lafayette are even larger than the average gap between queries and bookings. Here our hypothesis is that the gap between localness of queries, and bookings for a particular city might be correlated with the city's demographic information such as population and economy, plus impacted by geographic constraints (e.g., Los Angeles is on the ocean, whereas West Lafayette is in the middle of the country).

4.4 Summary

So far, we have modeled the spatial preference of users across both datasets and measured the relative geo-spatial "footprint" of different locations via their mean contribution distance, the radius of gyration, and the city locality. We have observed that the private queries (and bookings) are less local than the public check-ins, which casts doubt on the possibility of publicly shared location information serving as a substitute for privately held information. On an encouraging note, though, we have seen that the relative localness values are surprisingly consistent. Continuing this exploration of the spatial preference, we next turn to two studies designed to leverage spatial preference:

- In the first study, we develop a PageRank-like random walk for identifying spatially significant locations based on the spatial preference of users. Do we find that – in spite of their fundamental differences – that the two datasets reveal similar significant locations?
- In the second study, we investigate the potential of mining related clusters of locations from both datasets based on the spatial preferences of users. Do we find comparable performance across datasets? Or does one perform significantly better than the other?

5. STUDY 1: SPATIAL IMPACT

In this section, we explore the possibility of aggregating spatial preference information from multiple locations to provide a global perspective on the most "impactful" locations. Automatically deriving the significant locations from a location dataset is an important problem, and one that has potential applications in urban planning (e.g., what neighborhoods are highly-preferred and potentially facing an influx of new residents?), in location-based advertising (e.g., what points-of-interest are more important for a particular demographic target group?), among many others.

In the following, we formally define two approaches for extracting the significant locations from a location dataset and then we examine the locations identified over the private query dataset and the public check-in dataset.

5.1 Two Methods for Finding Spatial Impact

For a collection of locations $\mathscr{L}$, our goal is to find an ordering over the locations in $\mathscr{L}$ corresponding the relative spatial impact of locations, so that higher-ranked locations are deemed more signif-

Table 3: Examples of Impact Metrics

Impact Metric	ImpactRank			D-ImpactRank		
City Name	Queries	Bookings	Check-ins	Queries	Bookings	Check-ins
New York City	0.035931 (2)	0.017182 (2)	0.010677 (1)	0.038831 (2)	0.019854 (2)	0.016864 (1)
Los Angeles	0.013607 (9)	0.006141 (11)	0.005895 (7)	0.016910 (6)	0.008049 (8)	0.009194 (6)
Corpus Christi, TX	0.001476 (62)	0.001271 (89)	0.000458 (146)	0.001077 (74)	0.001019 (103)	0.000309 (206)
West Lafayette, IN	0.000073 (726)	0.000065 (1628)	0.000121 (535)	0.000064 (747)	0.000063 (1575)	0.000101 (534)

Table 4: Top 10 Most Impactful Cities By ImpactRank

	$Queries$	$Bookings$	$Check-ins$
No.1	Las Vegas	Las Vegas	New York
No.2	New York	New York	Austin
No.3	Orlando	Chicago	Orlando
No.4	Miami	Orlando	San Francisco
No.5	Chicago	San Diego	Las Vegas
No.6	San Francisco	Miami	Chicago
No.7	San Diego	New Orleans	Los Angeles
No.8	Phoenix	Washington, DC	Bay Lake, FL
No.9	Los Angeles	San Antonio	Anaheim
No.10	Washington, DC	Atlanta	Seattle

Table 5: Top 10 Most Impactful Cities By D-ImpactRank

	$Queries$	$Bookings$	$Check-ins$
No.1	Las Vegas	Las Vegas	New York
No.2	New York	New York	Austin
No.3	Orlando	San Francisco	San Francisco
No.4	San Francisco	Chicago	Orlando
No.5	Miami	San Diego	Las Vegas
No.6	Los Angeles	Seattle	Los Angeles
No.7	San Diego	Los Angeles	Chicago
No.8	Chicago	New Orleans	Seattle
No.9	New Orleans	Washington, DC	Bay Lake, FL
No.10	Washington, DC	Miami	Anaheim

icant than lower-ranked locations. While the notion of spatial impact is difficult to evaluate, we examine two approaches grounded in popular web link analysis and assess the orderings generated by each:

ImpactRank: The first approach propagates the spatial preference from one location to another, so that in aggregate the locations that are most preferred by locations that are themselves highly-preferred are the most "impactful". Similar to the PageRank approach for aggregating web links to assign a global importance score to web pages, ImpactRank can be viewed from the perspective of a biased random walker. At each location, the random walker chooses to visit a subsequent location based on the spatial preference of the current location. As in PageRank, the random walker occasionally loses interest in his travels and randomly picks a new starting location. In the limit, this random walk results in a global ordering over all locations based on the time spent by the random walker in each location.

Let S be the set of all locations, and let $S(\rightarrow l_i)$ be the set of all locations that express a non-negative spatial preference in l_i, such that $p(l_j, l_i)$ is the spatial preference probability of l_j toward l_i. The ImpactRank for location l_i, denoted by $IR(l_i)$, is then given by:

$$IR(l_i) = d \sum_{l_j \in S(l_i)} IR(l_j) p(l_j, l_i) + (1-d) \frac{1}{|S|}$$

where d is a damping factor (fixed as 0.85 in our experiments). The ImpactRank scores may be updated iteratively using the power method.

D-ImpactRank: ImpactRank measures the impact of a particular location purely based on the spatial preference matrix (which is essentially a transition matrix defined over locations), but without consideration for the actual distance between locations. Our goal is to incorporate this distance so that more distant locations are more rewarded for the same degree of spatial preference than closer locations. For example, suppose the spatial preference from A to B is 0.2 and from A to C is 0.2. If A and B are neighboring cities, but A and C are separated by 100s of miles, then this method can reward city C more since it has attracted interest from farther away. Thus, we extend ImpactRank to D-ImpactRank, by incorporating the physical distance between locations.

Specifically, we calculate the mean contribution distance (MCD) between all pairs of locations. Then for each spatial preference probability from an origin l_i to a destination l_j, we multiply the original probability by a weight of the distance between l_i and l_j divided by the weighted average distance. The distance weighted spatial preference probability $p'(l_j, l_i)$ from l_j to l_i is defined as:

$$p'(l_j, l_i) = p(l_j, l_i) * \frac{dist(l_j, l_i)}{MCD}$$

Then the D-ImpactRank scores are calculated with the distance weighted spatial preference matrix, and the D-ImpactRank scores for cities are expected to reveal both the cities' spatial impacts and the distance of their impacts' reach. The D-ImpactRank for location l_i can then be defined as in ImpactRank but with updated transition probabilities:

$$DIR(l_i) = d \sum_{l_j \in S(l_i)} IR(l_j) p'(l_j, l_i) + (1-d) \frac{1}{|S|}$$

5.2 Measuring Impact

Given the two approaches for measuring spatial impact, we calculate both over the private queries (and bookings) and the public check-ins. We apply each method to the cities in the Continental United States with dense spatial preference data. As before, we filter cities with fewer than 1000 queries (or bookings) and cities with fewer than 1000 check-ins.

We begin by continuing with our earlier example cities – New York City, Los Angeles, Corpus Christi, and West Lafayette – and listing their spatial impact scores and ranks (in parentheses) in Table 3. The relative rankings across both approaches and across all three datasets are remarkably consistent with New York > Los Angeles > Corpus Christi > West Lafayette. This is an encouraging result and one that fits well with our intuition (especially considering that Corpus Christi is a popular regional tourist destination as compared with the college town of West Lafayette).

We next list the top-10 cities with the highest spatial impact in Table 4 and Table 5, again considering both approaches and all three datasets. Focusing on Table 4, we see that five of the ten cities are common between the public check-in dataset and the private query dataset: New York, Orlando, San Francisco, Chicago, and Los Angeles. Note that Austin is the original home of the Gowalla location sharing service and so it receives a large "home field advantage". Bay Lake, Florida is the home of Walt Disney World next to Orlando and so could be considered a sixth similar location across the public and private datasets. Similarly, we see in Table 5 comparable rankings for the distance-weighted D-ImpactRank with

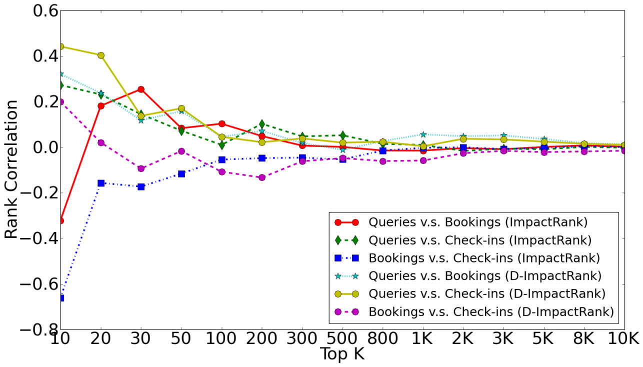


Figure 4: (Color) Rank Correlation between List of Most Impactful Cities

respect to the original ImpactRank.

Comparing between ImpactRank and D-ImpactRank for only the top-10 reveals little difference. Hence, we next measure the rank correlation across approaches using Spearman's ρ, which ranges from 1 to -1, with higher values indicating that two ranked lists are in relative agreement. As we can see in Figure 4, the rank correlation between approaches and between different datasets varies quite a bit. The series of red, green, and blue indicate the rank correlations between lists of top-K most impactful cities ranked by their ImpactRank scores. The series of cyan, yellow, and magenta indicate the rank correlations between lists of top K most impactful cities ranked by their D-ImpactRank scores. We are encouraged to see that the rank correlation for D-ImpactRank for queries versus check-ins performs very well over the top-20 results (meaning that the top-20 are highly correlated based on these two datasets). For bookings versus check-ins over ImpactRank (in blue), the rank correlation is the worst for K up to 100. At higher values of K, the rank correlation in all cases converges to around 0.0 primarily due to data sparsity at the bottom of the ranked list (leading to essentially random rankings at the bottom of the list).

Based on this experimental study, we find that in some cases both datasets reveal similar significant locations. This result is somewhat surprising considering the key differences between the public check-ins and the private queries, but is encouraging. In our following study, we continue this exploration of the viability of substituting publicly-released data for private data with an examination of extracting similar cities from location datasets.

6. STUDY 2: FINDING SIMILAR CITIES

In previous sections, we characterized a location by its spatial preference and by the spatial impact derived from aggregating over these spatial preferences. In this section, we examine whether these spatial characterizations can be used to automatically extract groups of similar locations. Finding related groups of locations has potential impact for optimizing online advertising (e.g., if users in location A click on an ad, then perhaps users in the similar location B will also do so), for improving web search and mobile applications (e.g., a user querying for a nearby tourist destination can be recommended other similar spots), and so forth.

Toward finding similar cities, we first define a ground truth of city similarity, define two metrics for evaluating city similarity, and then measure city similarity using a vector space interpretation of spatial preference and spatial impact.

6.1 Defining the Ground Truth

What makes two locations similar? While there are many possible answers, we adopt a systematic method for finding relationships among cities by mining 800 expert-curated lists of top cities across particular categories. The data is available from [1] and lists 101 top cities for each category. For example one of the lists includes the top cities with the most people having a Doctorate degree; for this list the top cities are Palo Alto (CA), Bethesda (MD), Brookline (MA), Cambridge (MA), and Davis (CA). From this perspective, these five cities can be considered similar. In this same fashion, we extract the top cities lists for a total of 800 separate city lists. For each pair of cities, we consider their total number of co-occurrences among the top city lists as the similarity between the pair of cities. For example, if two cities co-occur in 400 out of the 800 lists, then their similarity is $\frac{1}{2}$. Cities that never co-occur on a list will have a similarity of 0. In addition, for city l_i, we rank the other cities according to their similarities (co-occurrences in top city lists) with city l_i in descending order.

A similar approach was undertaken in the context of free-text web search engine queries in [15]. Rather than considering spatial preference as in this paper, the authors looked for common clues in the text of search engine queries to group related cities. Information revealed through text queries is a strong indicator of similarity (e.g., if many users in two locations are both querying for "molecular biology", "PhD", and "grad school", then there is good evidence of a relationship between locations). In contrast, spatial preference is a less clear indicator of city similarity since only relative interest in other locations is available for comparison.

6.2 Approach and Metrics

To find related cities, we apply the standard cosine similarity to vectors based on the spatial preference and the spatial impact of city pairs. That is, for city i and city j, we can represent each city by a vector (e.g., based on the spatial preference probabilities). Cosine similarity is a similarity measurement between the two vectors – in this case, the vectors associated with city i, $\vec{v}_i$, and with city j, $\vec{v}_j$:

$$cos(\vec{v}_i, \vec{v}_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i|\,|\vec{v}_j|}$$

With this approach and the ground-truth data, we use **Average Precision@10** ($P@10$) and **Average NDCG@10** ($N@10$) to evaluate the predicted top similar cities. For each city, we first extract the top K% of the most similar cities to it in the ground truth data as the relevant cities to the given city. Then, we calculate the Precision@10 for the city which measures the percentage of the top 10 predicted similar cities that also belong to the top K% of the relevant set, which can be formally defined as:

$$P@10 = \frac{\sum_{l_i \in S} \frac{|S_{top10}(l_i) \cap S_{top\_k\%\_gt}(l_i)|}{10}}{|S_c|}$$

where S refers to the set of all the cities in the datasets; l_i denotes a specific city; $S_{top10}(l_i)$ denotes the top 10 similar cities of l_i predicted using the similarity metric; and $S_{top\_k\%\_gt}(l_i)$ denotes the top K% similar cities for l_i in the ground-truth data.

A high value of $AvgPrecision@10$ indicates that the location preferences or localness modeled from the data really reveal semantic information for the city, and hence provide hints to find similar cities. Similarly, we apply **Average NDCG@10** to evaluate the performance considering both the precision of the predicted similar cities and the positions of the truly similar cities in the predicted similar city list.

In practice, we extract 10% of the ground truth similar cities for each city as its ground truth relevant cities. To make sure we have dense data for each of the cities, for both the queries and bookings, we only pick the cities in Continental United States with a minimum of 5000 queries from each of the city; and for the check-ins,

Table 6: Performance for Identifying Similar Cities

	Queries		Bookings		Check-ins	
Feature Set	P@10	N@10	P@10	N@10	P@10	N@10
Spatial Preference	25.17%	56.11%	28.06%	59.81%	24.2%	60.1%
Spatial Impact	22.22%	50.43%	24.71%	52.86%	31.2%	64.7%
Spatial Preference + Impact	28.04%	59.42%	28.97%	60.38%	31.6%	65.2%

we only pick the cities in Continental United States with a minimum of 1000 check-ins.

6.3 Evaluation

Table 6 shows the performance using features of different combinations of spatial preference and spatial impact associated with the private queries (and bookings) and the public check-ins. We additionally consider a combined vector representation that is simply an average of the normalized spatial preference and the normalized spatial impact vectors. Using cosine similarity to calculate the similarity between these three representations of cities, we observe strikingly similar results across the public check-ins and the private queries, as well as fairly stable relative ordering with the combined representation always yielding the best results.

Focusing on precision@10, we see that about 28% of the top-10 predicted similar cities are considered similar (based on the ground truth data) based on the query data, but that about 32% are similar based on the check-in data. Focusing on the average NDCG@10, we see a similar behavior – with the query data yielding a 60% result, but the check-ins performing slightly better with 65%.

Based on this experimental study, we find that across these two fundamentally different datasets, that similar performance may be achieved for automatically identifying groups of related locations. Coupled with the observations in the previous section, this is a second encouraging result considering the key differences between the two datasets.

7. CONCLUSION

In this paper, we have investigated two different sources of spatial preference: a set of private query logs recorded by a commercial hotel search engine and a set of publicly shared check-ins voluntarily generated by users of a popular location sharing service. Although generated by users with fundamentally different intentions, we find common conclusions may be drawn from both data sources, indicating the viability of publicly shared location information to complement (and replace, in some cases), privately held location information. This is especially encouraging since many location preference data sources are expensive, proprietary, and often times unavailable. In contrast, publicly-shared data offers appealing new avenues of research. Since modeling and exploiting spatial preference is critical for geographers, social scientists, as well as computer scientists interested in improving location-based recommendation systems, travel planners, search engines, and other emerging mobile applications, these conclusions are a starting point for further research on the strengths and weaknesses of relying on publicly available datasets.

Acknowledgements

This work was supported in part by NSF grant IIS-1149383. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] I. Advameg. Profiles of all u.s. cities, 2008. http://www.city-data.com.

[2] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW*, 2008.

[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, 2010.

[4] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[5] Z. Cheng, J. Caverlee, and K. Lee. Exploring millions of footprints in location sharing services. In *ICWSM*, 2011.

[6] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, 2011.

[7] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, 2008.

[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[9] K. E. Haynes and A. Fotheringham. *Gravity and Spatial Interaction Models*. Sage-Publications, 1984.

[10] B. Hecht and D. Gergle. On the "localness" of user-generated content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 2010.

[11] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi. Geographic dissection of the twitter network. In *ICWSM*, 2012.

[12] J. McGee, J. A. Caverlee, and Z. Cheng. A geographic study of tie strength in social media. In *CIKM*, 2011.

[13] R. Saga, Y. Hayashi, and H. Tsuji. Hotel recommender system based on user's preference transition. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, oct. 2008.

[14] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance matters: geo-social metrics for online social networks. In *Proceedings of the 3rd conference on Online social networks*, 2010.

[15] R. Seth, M. Covell, D. Ravichandran, D. Sivakumar, and S. Baluja. A tale of two (similar) cities: Inferring city similarity through geo-spatial query log analysis. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2011.

[16] B. Shaw. Machine learning with large networks of people and places, 2012. http://engineering.foursquare.com/2012/03/23/machine-learning-with-large-networks-of-people-and-places/.

[17] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[18] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*, 2011.

[19] Y. Zheng et al. Mining interesting locations and travel sequences from GPS trajectories. In *WWW '10*, 2009.