Querying and Exploring Polygamous Relationships in Urban Spatio-Temporal Data Sets

Yeuk-Yin Chan
New York University
yyc337@nyu.edu

Fernando Chirigati
New York University
fchirigati@nyu.edu

Harish Doraiswamy
New York University
harishd@nyu.edu

Cláudio T. Silva
New York University
csilva@nyu.edu

Juliana Freire
New York University
juliana.freire@nyu.edu

ABSTRACT

We present a system designed for exploring urban data set relationships introduced by the Data Polygamy framework, which are useful to uncover interesting patterns and interactions between the different components of a city. Since there can be a plethora of relationships to analyze, our interface helps discover relationships that are potentially interesting by allowing users to visually query and explore the relationship set. We will demonstrate the effectiveness of such interface through a few case studies, and demo visitors will be also able to do their own exploration.

KEYWORDS

Data polygamy, data set relationships, urban data

1 INTRODUCTION

Our increasing ability to collect, transmit, and store data, coupled with the growing trend towards openness, creates a unique opportunity to enable cities to deliver services efficiently and sustainably while keeping their citizens safe and well-informed. The challenge now lies in making sense of all the available data so that they can be used effectively by city agencies.

Urban data is unique in that it captures the behavior of the different components of a city, namely its residents, existing infrastructure (physical and policies), and the environment. While exploring a city's data exhaust to study such components, an expert may find an unexpected pattern or feature in a data set that may be explained by other related data. For example, consider the top plot in Figure 1, which shows the number of daily taxi trips in New York City (NYC) in 2011. While the distribution of trips tends to follow a pattern over time, we observe some atypical drops in August. A natural question is what might have caused these drastic reductions. By examining precipitation data (bottom plot in Figure 1), we discover that these drops occur on days with unusually high precipitation levels: the first two peaks are related to unexpected heavy rainfalls that disrupted traffic in the city,[1] and the last one was caused by hurricane Irene.

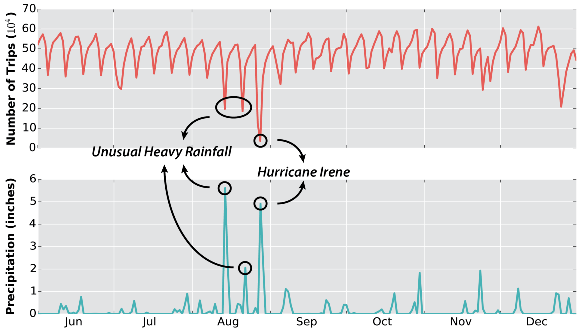


Figure 1: Variation of the number of taxi trips in NYC and its relationship with precipitation.

Therefore, in order to understand a city, it is important for an expert to also understand *how the different components interact over space and time*. The discovery and analysis of potential *relationships* between disparate data can lead to new hypotheses that explain phenomena represented in the data. In the previous example, the hypothesis would be that atypically heavy rainfall leads to significant reduction in the number of taxi trips. In addition to enabling *hypothesis generation*, studying relationships among data sets can also help with *hypothesis testing*, helping experts frame appropriate policies to counter them.

Uncovering relationships between disparate urban data sets is challenging for many reasons. First, urban data can be large vertically, containing hundreds of millions to billions of data points, and horizontally, consisting of several attributes [1]. Second, there is a large number of urban data sets: NYC alone has published over 1,300 data sets in the past two years [3], and this is just a small fraction of the data collected by the city. Also, urban data contain both spatial and temporal attributes at different resolutions, which requires an expert to look for data interactions in multiple resolutions. Because a relationship can exist between any pair of attributes, the data complexity coupled with the sheer number of available data sets and attributes creates a combinatorially large number of possible relationships where only a small fraction is potentially meaningful.

To address these challenges, we proposed Data Polygamy [2], a framework that efficiently identifies meaningful relationships across

SIGMOD'17, Raleigh, NC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

[1] http://nyti.ms/2inOVJH

urban data sets. We introduced the notion of *topology-based relationships*, where two data sets are related if there is a relationship between the *salient features* of the data (e.g., the atypical drops and peaks in Figure 1). Users can then pose *relationship queries*: hypothesis generation is supported by querying for relationships among all data sets, and a hypothesis can be tested by querying for relationships between the data sets involved in the hypothesis. To the best of our knowledge, no existing method addressed the problem of efficiently identifying spatio-temporal relationships that take into account salient features in the data.

Given a collection of data sets such as NYC Open Data [3], even after pruning relationships that are not statistically significant [2], users are still left with many relationships to analyze, both to discover potentially interesting relationships as well as to assess their validity. In this demo, we present a visual interface to assist users query and explore relationships. Since the many-to-many polygamous relationships imply a graph, we designed an interface that allows users to inspect this graph and interactively query the relationship set. Besides describing the system and its design, we present case studies that demonstrate its effectiveness.

2 THE DATA POLYGAMY FRAMEWORK

The data polygamy framework primarily consists of two components: *feature computation*, and *relationship evaluation*, which uses the computed features to identify the relationships. In this section, we give a brief overview of our framework. For more details, see [2].

2.1 Feature Computation

Consider the precipitation data and the NYC taxi trip data depicted in Figure 1. There is no apparent relation between the two data during the normal course of time: it is only when the precipitation is unusually high that there is a connection with taxi trips. This is common among urban data sets, where relationships become visible only at spatio-temporal regions (locations in space and time) that behave differently compared to the regions' neighborhood. In the *Data Polygamy* framework, we explored the use of computational topology to identify these interesting relationships. In particular, we introduced the notion of *topology-based relationships*, where two data set attributes are related if there is a relationship between the *salient features* of the data.

To give some intuition for why and how we apply topology, suppose we model a time step in an urban data set as a terrain, where the height of each point of the terrain represents the data value at that spatial location. The terrain (or data) is mathematically modeled as a *scalar function*, which maps each point on the spatial domain to a real value representing the function value. Given a scalar function, the variation over space is captured by the peaks and valleys of this terrain. This can be extended to include time by modeling the data as a high dimensional terrain. The salient features would then correspond to the regions behaving differently from their neighborhood, and are inherently represented as tall peaks and deep valleys. For instance, Figure 2 depicts a terrain represented as scalar function f; given the threshold f_1, the red peaks correspond to the salient features of this function.

We use and extend efficient algorithms from computational topology to compute salient features in the framework. Such algorithms are generic, in the sense that they work on data having different dimensions and resolutions without requiring any modification. Features can also have arbitrary spatial structures and straddle multiple time intervals. We also designed a *data-driven strategy* to identify the required feature thresholds, since manually selecting them might not be feasible over all data sets, especially when there are hundreds of data sets each having several attributes.

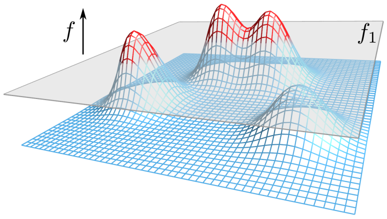


Figure 2: Example of a scalar function f represented as a terrain. The red peaks above the given threshold value f_1 correspond to the salient features of this function.

2.2 Relationship Evaluation

Given two attributes of a pair of data sets, to determine whether they are related, we assess how similar their corresponding terrains are, i.e., the similarities in the spatio-temporal variation patterns of the scalar functions representing the data. Thus, possible relationships are identified depending on the commonality between the salient features: two features are considered to be *related* if they occur at the same spatio-temporal region; they are *positively related* if both features are positive or negative, and *negatively related* otherwise (e.g., features from precipitation and NYC taxi trips corresponding to hurricane Irene in Figure 1 are negatively related). Relationships are then evaluated based on two measures: *score* and *strength*.

Relationship Score τ. This measure captures the overall nature (polarity) of the relationship, i.e., whether it is always positive (features are always positively related), always negative (features are always negatively related), or somewhere in between. τ ranges from -1 to 1: a value closer to -1 indicates the two attributes are negatively related, while a value closer to 1 indicates they are positively related.

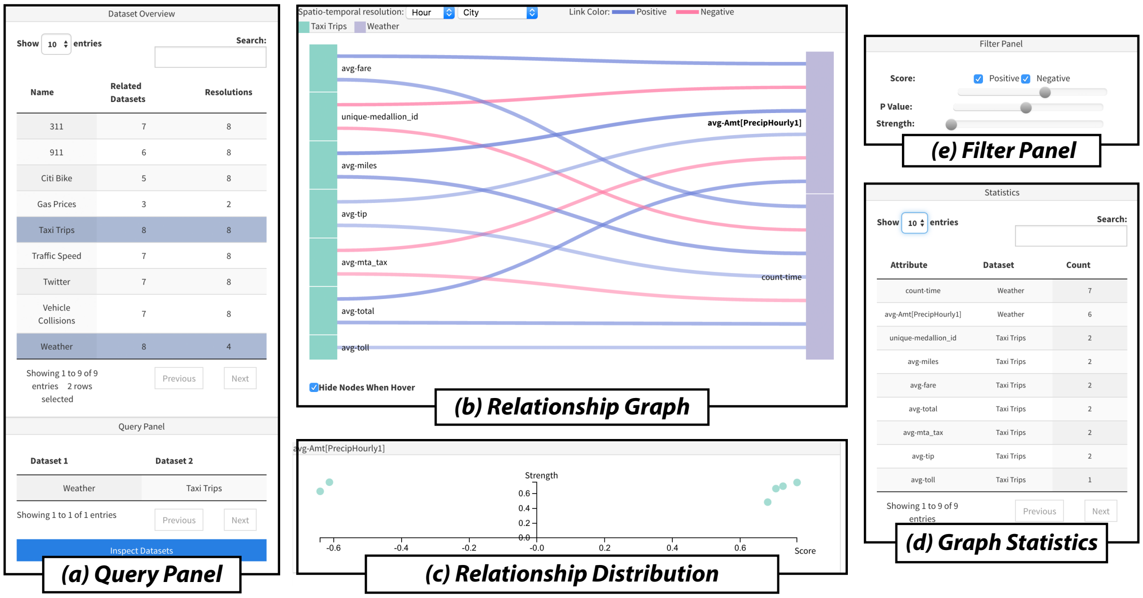
Relationship Strength ρ. This measure is used to capture how frequently features in the two attributes are related: the more frequently the features are related, the stronger the relationship is. ρ ranges from 0 to 1: a value of ρ closer to 1 indicates a strong relationship between the two attributes, since a feature in one attribute almost always indicates a feature in the other attribute as well. Similarly, a value of ρ close to 0 indicates a weak relationship.

In addition, the statistical significance for each relationship is also computed, and the corresponding ***p*-values** are returned. Since Monte Carlo methods assume independence across samples, to assess the statistical significance of a potential relationship, we developed restricted Monte Carlo permutation tests that respect data dependencies due to spatial and temporal proximity.

2.3 Relationship Querying

The framework supports the following *relationship query*:

Find relationships between $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *satisfying* clause



Dataset Overview

Show 10 entries Search:

Name	Related Datasets	Resolutions
311	7	8
911	6	8
Citi Bike	5	8
Gas Prices	3	2
Taxi Trips	8	8
Traffic Speed	7	8
Twitter	7	8
Vehicle Collisions	7	8
Weather	8	4

Showing 1 to 9 of 9 entries 2 rows selected

Query Panel

Dataset 1	Dataset 2
Weather	Taxi Trips

Showing 1 to 1 of 1 entries

Spatio-temporal resolution: Hour City

☑ Hide Nodes When Hover

Filter Panel

Score: ☑ Positive ☑ Negative

P Value:

Strength:

Statistics

Show 10 entries Search:

Attribute	Dataset	Count
count-time	Weather	7
avg-Amt[PrecipHourly1]	Weather	6
unique-medallion_id	Taxi Trips	2
avg-miles	Taxi Trips	2
avg-fare	Taxi Trips	2
avg-total	Taxi Trips	2
avg-mta_tax	Taxi Trips	2
avg-tip	Taxi Trips	2
avg-toll	Taxi Trips	1

Showing 1 to 9 of 9 entries

Figure 3: The user interface for the system and its different components.

where $\mathcal{D}_1$ and $\mathcal{D}_2$ are collections of data sets. Relationships across attributes of these two collections over all possible spatio-temporal resolutions are evaluated. In CLAUSE, optional condition parameters can be specified to filter relationships satisfying a minimum score τ or strength ρ, or a maximum p-value.

3 VISUALLY EXPLORING RELATIONSHIPS

Even after pruning relationships that are not statistically significant, the size of the relationship set can be overwhelming for users. As an example, a collection of 9 data sets can still generate as many as 100 relationships that are potentially meaningful for a single spatio-temporal resolution, and when working with the NYC Open Data, over 10,000 relationships are derived [2]. Our goal is to help users visually explore results from such relationship queries, thus assisting them in their analyses. To do this, we designed an interface that allows users to visually specify queries (mentioned in Section 2.3), as well as inspect and filter the results in an interactive manner.

3.1 Data Set Exploration

Users can select data sets for the query based on their main interest using the Query Panel (Figure 3(a)). This panel lists the available data sets and provides a relationship overview of each of these data sets to the user. In particular, for each data set $\mathbb{D}$, the following two properties are also shown: the number of data sets having at least one attribute relationship with $\mathbb{D}$; and the total number of spatio-temporal resolutions where $\mathbb{D}$ has at least one relationship. This helps users understand the extent of the polygamous relationships and the scope of the resolutions for different data sets.

One or more data sets are selected from the query panel in order to perform the query. If a single data set $\mathbb{D}$ is selected (the pivot data set), a *one-to-many query* is performed, i.e., the relationships between $\mathbb{D}$ and the remaining data sets are retrieved; if multiple data sets are selected (a collection $\mathcal{D}$), a *many-to-many query* is performed, i.e., all the relationships among the data sets in $\mathcal{D}$ are retrieved. For instance, in Figure 3(a), the user is interested in a *many-to-many query* between Taxi Trips and Weather data sets. Users can then fine-tune the query by filtering based on the different parameters, namely the score, strength, and p-value (see Figure 3(e)). Note that these are the same types of queries supported in the Data Polygamy framework [2].

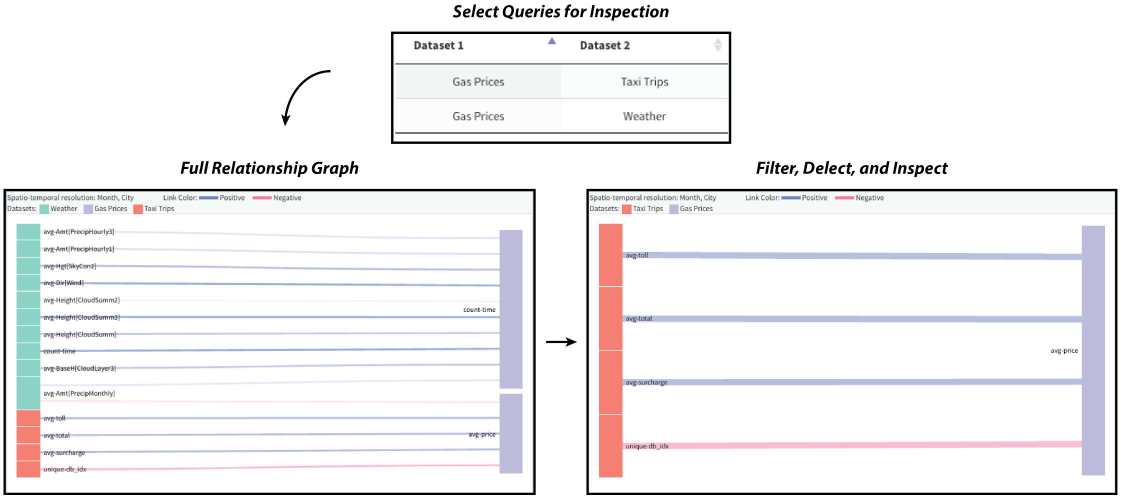
To further focus their analyses, the user can also select the spatio-temporal resolution they want to explore by using a combo box selection, which provides a list of all the available resolutions supported by the selected data sets. For example, in Figure 3(b), the desired resolutions are the city resolution with respect to space and the hourly resolution with respect to time—henceforth represented as (city, hour) resolution.

3.2 Relationship Exploration

The specified query results in a set of relationships, which, as mentioned earlier, can be prohibitively large. We therefore allow the visual inspection of these relationships, as well as provide visual cues to ease the analysis process.

Relationship Graph. Note that the set of relationships can be modeled as a graph, where the nodes correspond to attributes of the data sets, and edges denote a relationship. We use an interactive Sankey diagram, which computes a fully automatic layout in a force-directed approach to display these relationships between attributes in the most visible manner [4].

All nodes corresponding to the attributes of a given data set share the same color. The size of the node in this diagram encodes the number of relationships. The opacity of the edges representing the relationship encodes the strength (ρ) of the relationship. The score (τ) is encoded using the stroke width of the edge, while the polarity of the relationship is encoded using color (blue and red for positive and negative relationships, respectively). These visual cues give a useful overview of how the attributes are related. For instance, Figure 3(b) depicts the relationships between Taxi Trips and Weather



Select Queries for Inspection

Dataset 1	Dataset 2
Gas Prices	Taxi Trips
Gas Prices	Weather

Full Relationship Graph

Filter, Delect, and Inspect

Figure 4: Selecting relationships that can show the potential influence of a gas price increase.

data sets, and by looking at the graph, it is clear that most relationships are positive and strong. By hovering over the relationships, their properties are displayed, such as score and strength. To help with the exploration, users are also allowed to: re-arrange the position of the attributes by dragging them in the canvas; zoom in and out; select attributes of interest to focus the analysis on their relationships; and hide attributes that are not interesting.

As mentioned earlier, users can also interactively select the optional condition parameters of the query: score, strength, and p-value (Figure 3(e)). Moving the sliders or typing the thresholds dynamically updates the graph to reflect the remaining relationships after filtering. Since this filtering is interactive, it streamlines the exploration and helps users choose parameters that best suit their analyses.

Exploring Relationship Statistics. Overall statistics regarding the graph are provided (Figure 3(d)), in particular the number of relationships for each of the attributes. Whenever an attribute is selected in the graph, the distribution of the strength and score of the relationships involving that attribute is visualized as a scatter plot (Figure 3(c)): each point in the plot corresponds to a relationship, and the x- and y-axis of the plot correspond to the score and strength, respectively. Hovering over the points shows their properties. Users can hide relationships that are not of their interest by double-clicking on the corresponding points. The statistics and the distributions are automatically updated whenever the graph changes.

4 DEMONSTRATION

In our demonstration, we will allow visitors to interact with the system and visualize the relationships across a plethora of urban data sets. In addition to doing their own explorations, we will present a few case studies that show the usefulness of the interface, such as:

Why it is so hard to find a taxi when it is raining?
In NYC, residents have the impression that it is hard to find a taxi in rainy days. To test such hypothesis, and to generate one for the explanation, an expert can select the Taxi Trips and Weather data sets. When exploring the (city, hour) resolution, she can use the filter queries to focus on more meaningful relationships (e.g., $\tau > 0.6$ and p-value < 0.05) and remove all the Weather attributes except for the precipitation one. A (very strong) negative relationship between number of taxi medallions and precipitation can then be found, confirming the hypothesis. A positive relationship between average taxi fare and precipitation generates a hypothesis for the explanation: taxi drivers are target earners, and they stop working after reaching their goal [2]. This case study is depicted by Figure 3.

Would a reduction in traffic speed reduce fatalities?
To test such hypothesis, first, an expert can select both Traffic Speed and Vehicle Collisions data sets from the Query Panel. By choosing the (neighborhood, day) resolution, and retaining only the relationships with respect to the speed attribute from the Traffic Speed data set, many positive relationships can be found between speed and the number on fatalities. By inspecting relationships from the distribution, the expert can also see that they all have a high score, which indicates that the hypothesis is true, and that a reduction on the speed limit in the streets could be beneficial.

What influence does an increase in gas prices provide?
An expert starts by selecting the Gas Prices data set, and choosing, say, the (city, month) resolution. She then selects all the data sets that have relationships with the Gas Prices: Weather and Taxi Trips. By retaining only the average gas price attribute in the graph, a significant negative relationship between number of taxis and gas price can be found. This indicates that gas price increases influence on the number of taxis running in the city, which may create demand issues. This case study is depicted in Figure 4.

REFERENCES

[1] Luciano Barbosa, Kien Pham, Cláudio Silva, Marcos Vieira, and Juliana Freire. 2014. Structured Open Urban Data: Understanding the Landscape. *Big Data* 2, 3 (2014).

[2] Fernando Chirigati, Harish Doraiswamy, Theodoros Damoulas, and Juliana Freire. 2016. Data Polygamy: The Many-Many Relationships Among Urban Spatio-Temporal Data Sets. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16)*. 1011–1025.

[3] NYC Open Data. 2017. NYC Open Data. https://nycopendata.socrata.com. (2017). Accessed: 2017-01-15.

[4] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. 2005. Interactive Sankey Diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*. IEEE, 233–240.