Analyzing Social Media to Characterize Local HIV At-risk Populations

Narendran Thangarajan
Computer Science and Engineering
UC San Diego
La Jolla, CA, USA
naren@eng.ucsd.edu

Nella Green
Department of Medicine
Antiviral Research Center
UC San Diego
La Jolla, CA, USA
n2green@ucsd.edu

Amarnath Gupta
San Diego Supercomputer Center
UC San Diego
La Jolla, CA, USA
a1gupta@ucsd.edu

Susan Little
Department of Medicine
Antiviral Research Center
UC San Diego
La Jolla, CA, USA
slittle@ucsd.edu

Nadir Weibel
Computer Science and Engineering
UC San Diego
La Jolla, CA, USA
weibel@ucsd.edu

ABSTRACT

The number of new HIV infections per year in the U.S. has remained stable at 50,000 since the 1990's. To improve epidemic control, we need more public health tools that are aimed at decreasing HIV transmission. Online social networks and their real-time communication capabilities are emerging as novel platforms for conducting epidemiological studies and recent research has outlined the feasibility of using Twitter to study HIV epidemiology. We propose a new method for identifying HIV at-risk populations using publicly available data from Twitter as an indicator of HIV risk. In this paper we take existing approaches further by introducing a new infrastructure to collect, classify, query and visualize these data, and we show the feasibility of identifying and characterizing HIV at-risk populations in the San Diego area at a finer level of granularity.

Categories and Subject Descriptors

H.1.m [**Information Systems**]: Models and Principles—*Miscellaneous*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query Formulation, Retrieval, Search Process*; H.2.1 [**Database Management**]: Logical Design—*Data Models*; H.5.m [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

General Terms

Design, Experimentation

Keywords

Twitter, Social Networks, Graph Modeling, Data Analysis, Visualizations, HIV, Prevention, Digital Epidemiology

Wireless Health 15 Oct 14-16, 2015, Bethesda, MD, USA
Copyright 2015 ACM 978-1-4503-3851-6 ...$15.00.
http://dx.doi.org/10.1145/2811780.2811923

1. INTRODUCTION

We can define social media as a group of digital tools enabling any user to easily publish and share information through the internet. The incredible impact of social media can be seen in the extent and growth of the social media population: currently 71% of the online adult population and 56% of the overall population in the U.S. use social media. In 2015, 176M people in America were using social media as compared to 163.6M people in 2011 [10, 25]. The U.S. are the 7th most connected country with 87% of the population using the internet and this trend is further accentuated by the rapid adoption of smart phones due to production of cheaper handsets and more affordable data plans [25].

Many recent studies showed how people's behavior on online social networks closely resembles their real life [24, 21], suggesting that the huge trove of data from these online social networks could provide an accurate description of people's day-to-day lives. As described below, it has been shown that there is a very good correlation between people's real and social life and we believe that this realization has the potential to open up myriads of possibilities in terms of public health epidemiology, as well as real-world interventions. Specifically, in this project we intend to exploit this valuable information to improve the quality of epidemic control programs focused on HIV. Currently, more than 1.2 million people in United States are infected with HIV. This count is increasing at the steady pace of 50,000 every year since 1990's, leading to discussions among public health workers about the need for more efficacious epidemic control strategies. For the purposes of this study, we consider this number, 50,000, as the null result. According to the HIV surveillance supplemental report from 2012, among the 50,000 new infections in the U.S. every year, 28,500 of them are found among men who have sex with men (MSM) [8]. Based on clinical data, the chances of acquiring HIV have been correlated to certain risky behaviors like drug consumption, unprotected sex, and others.

We based our approach and hypothesis on these data, and we posit that given how people's social media posts mirror (to a large extent) their personal life, it is possible to build an infection surveillance radar of HIV transmission risk behavior through real-time analysis of posts to online social networks. We decided to build our system on the Twitter infrastructure because of the publicly available data and because of the level of anonymity it provides to its users. This anonymity allows users to be part of communities they are really interested in, and post content which more accurately matches with real-life behaviors. Our ultimate goal is to learn how actual HIV at-risk users behave on Twitter and build a computational model that can then be applied to other Twitter users to investigate similarities in their behaviors through real-time visualizations, such as heat maps of particular geographical regions showing higher risk of HIV. In this paper we discuss the procedure to model the information obtained from Twitter so that they can be compared to real-world HIV transmission networks as the one curated by the AntiViral Research Center (AVRC) at UC San Diego [16]. The goal is to evaluate opportunities to target HIV testing and prevention efforts to communities at greatest risk of HIV acquisition, specifically in San Diego, California.

To achieve our goal, we built a computational infrastructure and information system based on data published on Twitter. Our system introduces an effective way to collect, clean, classify, organize and manage the stream of tweets that are constantly published online every second. Additionally in this paper we discuss our approach to accurately classify Twitter posts as "HIV at-risk" to minimize false positives. We introduce a way to model evolution of network entities like Twitter users and how their tweets' content, interaction and relationships with other Twitter users change with time. Finally we highlight our first results describing how HIV-related public health can benefit from digital epidemiology studies using Twitter.

In summary, our contributions through this paper are three-fold:

1. A system-level concept and design for using social media applications (like Twitter) to create a real-time radar of high risk areas based on behaviors related to infectious diseases. In our case, we focus on HIV transmission risk behaviors in San Diego, California.
2. A structured approach for collecting, classifying, cleaning, modeling, organizing and analyzing streaming social data such as tweets.
3. A series of insights from exploratory analysis of social network data to unravel latent patterns and hidden relationships between user behaviors. Specifically, we show how different HIV transmission risk behaviors are correlated among each other.

2. BACKGROUND AND RELATED WORK

State-of-the-art approaches to tracking the spread of infectious diseases such as SARS, influenza, HIV and others are typically based on *Syndromic Surveillance* techniques [14]. Syndromic surveillance involves analysis of medical data including Electronic Health Records (EHR) to detect and predict outbreaks of infectious diseases in a specific location. This methodology is based on the wide usage of digital tools for maintaining health records which speed up creation, editing, sharing and analysis of patient records. The first practical implementation of this kind of surveillance attempted to augment the traditional clinical data analysis approach with real-time information from users' online search behaviors. It was used for tracking influenza in 2004 and results were published in 2006 [12]. Further research in this direction involved the employment of machine learning and natural language processing (NLP) techniques to extract more relevant information from social networks which improved the overall accuracy of the procedure [2].

While syndromic surveillance has proven to be an interesting approach to understand the spread of infections [11, 19], *Digital Epidemiology* takes this methodology a step further by making use of digital data sources like mobile phones, social media and sensor data to help in detection, prediction and etiology of infections [23]. This novel trend initially leveraged online news channels and real-time sources of interesting events (related to health care) and was mainly focused around Brownstein's work at Harvard University starting in 2009 [7]. Then the trend shifted to search engine queries leading to a line of research trying to find patterns in online search engine queries [1, 9, 3]. Most recent work on digital epidemiology increasingly focuses on using platforms like Twitter and Facebook.

Given the novelty of digital epidemiology and the recent emergence of social network analysis, there are few published data on tracking and characterizing HIV infections using online social networks. In 2014, the public health officials in Milwaukee, Wisconsin published a report [6] detailing how they used Facebook to identify and reach out to partners of participants who are HIV-positive quickly and easily to perform targeted intervention. This effectively augmented the traditional HIV intervention approaches and accelerated the "partner notification" strategy. However, this study was focused on a smaller geographical area and was human resource-heavy, leading to long turnaround times from detection to intervention.

Researchers from UCLA recently studied Twitter data across the U.S. and showed convincing proof that social media was ready to be mined for epidemiological analysis of HIV [27]. Similar research was conducted at Microsoft Research to study the social behavior of people trying to quit smoking [18]. They concluded that the people who are successful in their cessation attempts have markedly distinct social network characteristics as compared to those who fail cessation. Fowler et. al. also published results relating social media with behaviors, showing for instance that a person's happiness can be influenced by their position in their social network [13]. These data, and others, show that people's real-life behaviors can be modeled from social media relationships and interactions.

In our study we focus on the spread of HIV in San Diego County and the detection of HIV risk at the level of individuals and communities. We collect publicly available tweets and classify them to inform tailored prevention efforts by building a model of HIV risk based on a variety of dimensions such as geography, demographics, and social groups. In the remainder of this paper we describe our approach towards data collection, data cleaning, and data classification, and we show how our infrastructure, based on a graph database, allows us to extract important information about HIV risk behavior in the San Diego area, in particular for populations of men who have sex with men (MSM), who are the most at risk for HIV infection. Though our methodology is generic to all segments of the population, we initially included only the MSM community using interaction and relationship patterns among mentions of terms related to homosexuality on Twitter.

3. DATA COLLECTION

We decided to perform analysis on Twitter mainly because of the availability of public data (tweets) and the level of anonymity it provides to its users, but most importantly we wanted to build on earlier results published by Sean et al. [27]. Well-known social networking sites like Twitter provide Application Programming Interfaces (APIs), to programmatically access their data instead of scraping their websites. In addition to standard APIs, Twitter's engineering team developed a Streaming API that provide third-party developers low latency access to Twitter's global stream of tweets data in real-time. There are three kinds of Streaming API, the *Sample hose*, the *Fire hose* and the *Filter hose*. The Twitter Streaming API creates a long-standing connection between the client and the server and streams the incoming tweets to the subscribing clients.

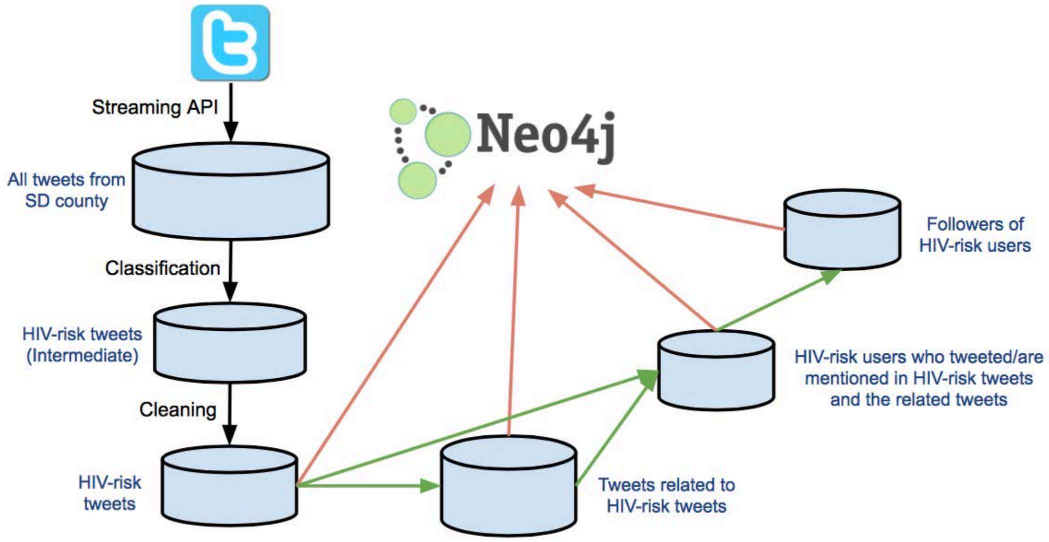
Streaming API
All tweets from SD county
Classification
HIV-risk tweets (Intermediate)
Cleaning
HIV-risk tweets
Tweets related to HIV-risk tweets
HIV-risk users who tweeted/are mentioned in HIV-risk tweets and the related tweets
Followers of HIV-risk users
Neo4j

Figure 1: Data Collection Architecture (i) Collection of tweets in real-time. (ii) Classification of tweets into HIV transmission risk tweets. (iii) Cleaning tweets to remove false positives. (iv) Deriving related tweets and users from HIV transmission risk tweets using Twitter's public API - green arrows (v) Construction of social network graph on Neo4j graph database - red arrows.

We started our project by consuming Twitter's sample hose API. The sample hose provides a 1% sample of global tweets at the rate of 70 tweets per second.

To narrow down tweets to the San Diego area, we needed to filter tweets based on their geographical origin. A "geotag" is a digital tag which encodes the location from where digital data originated. In tweets, it is encoded as a combination of longitude and latitude. Although it is required to know the geotag to understand if the tweet is of significance, research has shown that only less than 1.6% of the tweets from sample hose are geotagged [20]. Having less than 1.6% of this relatively small hose would have drastically lowered the amount of San Diego county data that we could collect. Hence, we moved to Twitter's filter hose API. The filter hose allows us to give a geocoded bounding box, and returns geo-tagged tweets that are generated within that geographical bounding box in real-time. So, in our case, we provided the geocoded bounding box for San Diego County. However, a restriction on filter hose is that the total number of tweets returned every second will not exceed 1% of the total number of tweets generated by Twitter during that second. Additionally, according to the study performed by researchers from Arizona State University and Carnegie Mellon University, in general only 1.45% of all the tweets are geotagged [17]. Even with the filter hose we could only consume a relatively low number of tweets from a specific geographical area. Despite these restrictions, we have collected more than 11.5 million tweets from the San Diego county region between October 2014 and August 2015.

3.1 Data Collection Architecture

The tweets collected using the Streaming API are pushed on to a mongoDB database that stores them and enables easy access for analysis. We choose mongoDB[1] as the data store mainly due to the ease of changing schema, high write throughput, and support for native map-reduce queries for performing on-demand aggregations. This model inherently exploits the parallelism in the current generation multiprocessor systems leading to high execution speeds. It is well suited for aggregation queries like word count, which in turn can power visualizations like word cloud or charts showing time-series information. To identify HIV at-risk tweets, we filter all the incoming tweets from the San Diego County and create a smaller corpus of tweets by classifying them based on the presence of certain HIV transmission risk words in the tweet's content. We exploited the domain expertise of our clinical collaborators to create five categories of HIV risk words and we filled those *buckets* with words commonly used in the local community in San Diego:

1. (Illicit) Drug Bucket
2. Sex Bucket
3. Bar/lounges/bath houses Bucket
4. Homosexual Terms Bucket
5. Sexually Transmitted Infections Bucket

Each bucket has around 15 HIV risk behavior terms; for example *meth* (Drug), *The Loft* (San Diego Bar) and *syphilis* (STI) providing a simple way to classify relevant terms in our tweets. After this phase, we perform further cleaning (see next section) to remove false positives and get a smaller subset of tweets that are related to HIV risk behaviors. By using the public Twitter APIs we then collect all related tweets like replies to the collected tweets and original tweets of re-tweets. We also pull from Twitter all the related users who either tweeted a HIV transmission risk tweet or were mentioned in a HIV transmission risk tweet. All these data are stored on separate mongoDB collections. Figure 1 illustrates our complete data collection infrastructure and overall process.

3.2 Privacy and Confidentiality

Social media research, as we can expect, involves privacy and ethical concerns related to the personally identifiable nature of the data collected and analyzed. In order to undertake this research, our protocols have been reviewed and approved by the Institutional Review Board of the University of California, San Diego. To prevent loss of confidentiality due to data breach, all of our data is de-identified and is stored on highly secure, password protected servers. Whenever a HIV-transmission risk user information is pulled from Twitter, the user ID is transformed using a one-way hash function and further manipulated to generate a new neutral user ID used solely for the purposes of the study. For all further experiments, we use this neutral user ID, making it impossible to map back results to the actual Twitter user. To provide further safety, all the results of this study are reported only in aggregate format.

[1] http://www.mongodb.org/

4. DATA CLEANING

As mentioned in Ren et. al. [22], Twitter is notorious for being inert to traditional NLP techniques mainly because of two properties. Firstly, the text is artificially limited to 140 characters encouraging twitter users to use emoticons and lots of text shortening and abbreviation tricks to workaround that limitation. Secondly, Twitter content being a social text stream is affected by concept drift, i.e. the same tweet could mean different things in different occasions. However, in our case, since we have a relatively small vocabulary of significant terms, we could effectively clean up the data based on presence/absence of phrases around these terms.

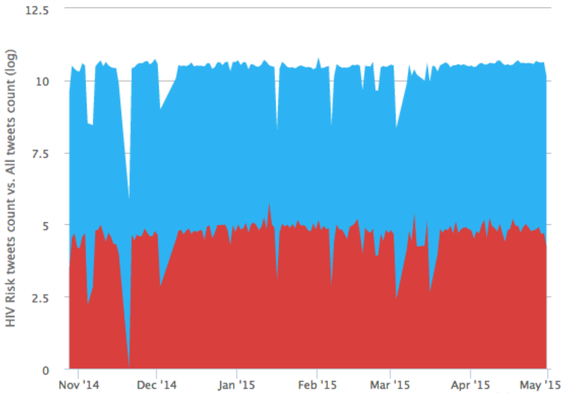

Figure 2: Trend of HIV transmission risk tweets per day between October 2014 and May 2015 (count values are in log). The blue area denotes all the tweets while the red area denotes the HIV transmission risk tweets alone.

As explained in the previous section, the data collection consists of filtering tweets that contained a pre-defined set of HIV risk words. We observed that for every risk word, we could find a set of words that can often co-occur with them and be indicative of whether the tweet as a whole exhibits HIV risk behavior or not. By manually scrutinizing around 1000 tweets, we created exclusion lists and inclusion lists for each HIV transmission risk word under every risk bucket. An example exclusion list for the HIV transmission risk word 'crack' (which falls under the risk bucket 'Drug') would include [crack me up, crack myself up, crack up, crack open, screen, joke, phone, system, hack, hacked]. In this exclusion list, the presence of the phrase 'crack me up' in the tweet indicates that the tweet is actually noisy and is not relevant to HIV risk behavior.

An example inclusion list in the risk bucket 'Bars/lounges/bath houses' for the risk word 'cheers' (a night club in San Diego) would include [san diego, sd, bar, drinks, drink, beer, @ cheers, @cheers]. We include "san diego" although all our geo-tagged tweets are from San Diego since online *checkins* from Cheers club shared through Twitter always have "San Diego" as part of the tweet. In this inclusion list, only presence of phrases like '@cheers', 'san diego', 'drinks' is indicative that the tweet is talking about being at the 'Cheers' club while every other tweet is probably not relevant in the context of the club. The data-cleaning phase led to reduction of HIV risk tweets by a significant 60%, resulting in a total of 35,000 risk tweets as of August 2015. Since the exclusion lists and inclusion lists are updated on a regular basis, based on new understanding about the particular domain, data cleaning runs as a batch process every 3 hours.

Only after the data-cleaning phase we start to understand the relationships between the tweets collected and their estimated relationship to HIV transmission risk. By looking at the data that are continuously collected we can now start to analyze the magnitude of HIV risk tweets. The chart shown in Fig. 2 shows the trend in the raw number of tweets exposing putative HIV risk behavior. Although our data showed a few periods of exceptionally high HIV transmission risk tweets related to unusual user behaviors, we calculated that an average of 15 HIV risk tweets are posted every hour.

Each tweet's location is inferred from the geo-coordinates embedded by the user's GPS-enabled device. A geo-coordinate from a GPS device, generally, has horizontal accuracy which can vary from 0.1 meter to 10 meters [15] which is good enough for our study. Using this information, we created a heat-map showing the geographic origin of HIV transmission risk tweets. The visualizations shown in Fig. 3 and Fig. 4 illustrate how the HIV risk tweets were distributed over space between October 2014 and May 2015. Note that this correlates with the overall number of tweets (risk and non-risk) and does not necessarily indicate that those regions are at higher risk. For instance, in Fig. 3, the maximum number of tweets that originated from a single location[2] is 11,009. Therefore, that particular spot becomes the reddest spot on the map and gradually blends into orange, yellow, green and blue as the number of tweets goes down. In Fig. 4, the red spot corresponds to 1.0, i.e. all the tweets are HIV transmission risk tweets. The facts that specific areas of the city have a very high number of tweets and that other places have relatively high *HIV transmission risk tweets to all tweets ratio* make them significant candidates for HIV risk areas.

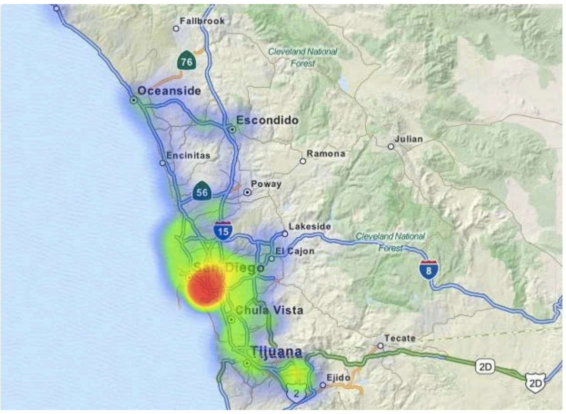

Figure 3: Heatmap showing all the tweets generated between October 2014 and May 2015 in the San Diego County area.

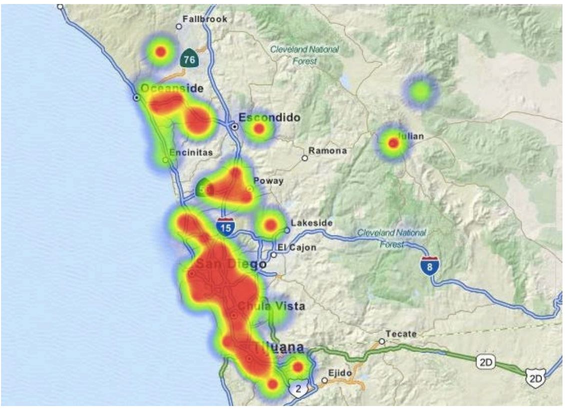

Figure 4: Heatmap showing ratio of HIV transmission risk tweets to all tweets generated 10/14 - 05/15 in San Diego.

[2] A "location" refers to a <lat,lon> value with 6 decimal precision.

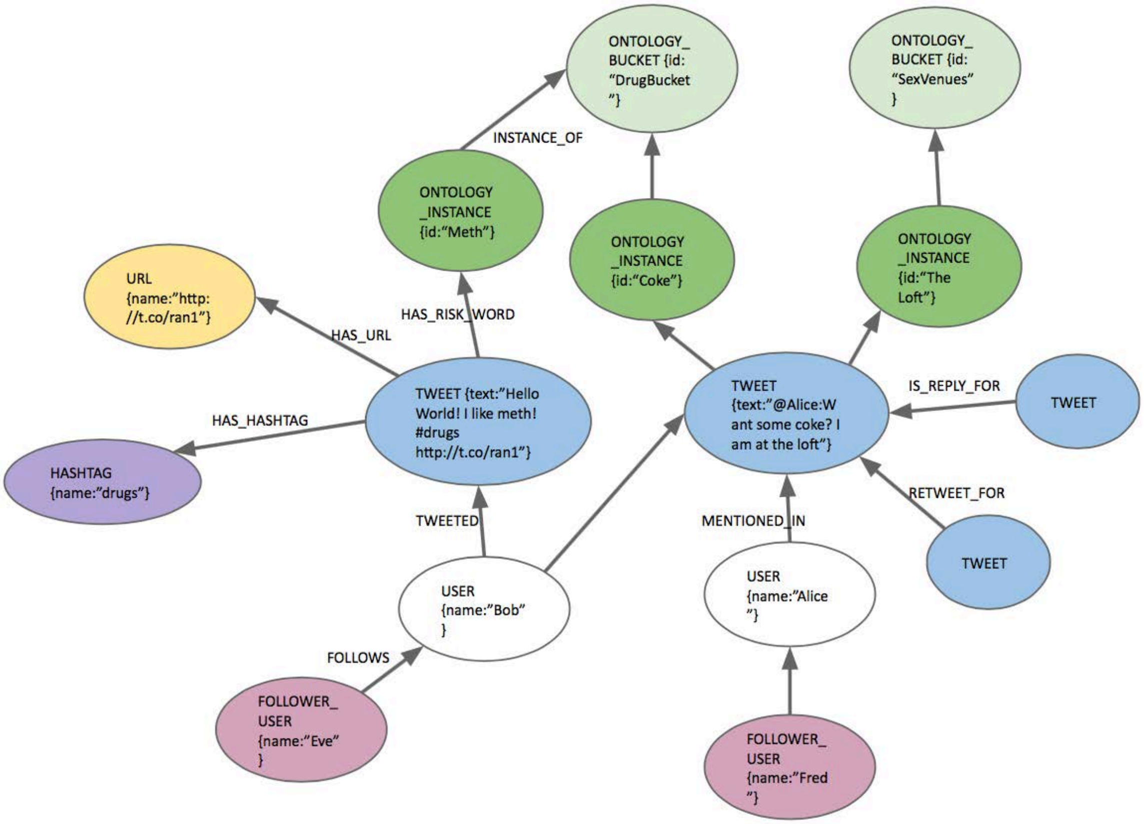
ONTOLOGY_ BUCKET {id: "DrugBucket "}
ONTOLOGY_ BUCKET {id: "SexVenues" }
INSTANCE_OF
ONTOLOGY _INSTANCE {id:"Meth"}
ONTOLOGY _INSTANCE {id:"Coke"}
ONTOLOGY _INSTANCE {id:"The Loft"}
URL {name:"http: //t.co/ran1"}
HAS_RISK_WORD
HAS_URL
TWEET {text:"Hello World! I like meth! #drugs http://t.co/ran1"}
TWEET {text:"@Alice:W ant some coke? I am at the loft"}
IS_REPLY_FOR
TWEET
HAS_HASHTAG
HASHTAG {name:"drugs"}
RETWEET_FOR
TWEETED
MENTIONED_IN
TWEET
USER {name:"Bob" }
USER {name:"Alice "}
FOLLOWS
FOLLOWER_ USER {name:"Eve" }
FOLLOWER_ USER {name:"Fred "}

Figure 5: Property Graph data model of the HIV at-risk social network

5. TWITTER RISK NETWORK MODEL

Besides raw analysis of magnitude and location of HIV risk tweets, we wanted to start reasoning about the relationships between users and tweets. The structure of Twitter and the available public APIs allow us to retrieve publicly available information about the characteristics of HIV at-risk users such as the information about the Twitter users they are following, if those users are mentioned in other tweets, or any hashtag (#) or URL used in their tweets. We modeled tweets and users in a graph around the concepts of nodes and relationships. During this process, we eliminated some Twitter fields like "contributors" which do not add value to the analysis we were planning to perform. Ultimately, our property graph data model, as shown in Fig. 5, is based on seven different types of nodes listed below.

- `USER` nodes correspond to each user in Twitter. Each `USER` node in our graph is an HIV risk Twitter user, classified based on his/her tweets.
- `TWEET` nodes corresponding to each HIV risk tweet in Twitter.
- `HASHTAG` nodes correspond to each hashtag used in the HIV risk tweets.
- `URL` nodes correspond to each URL being referred to in the HIV risk tweets.
- `FOLLOWER_USER` nodes correspond to the set of users that follow each of the `USER` nodes. So `FOLLOWER_USER` may or may not be an HIV risk Twitter user.
- `ONTOLOGY_BUCKET` nodes correspond to each of the five risk buckets we defined above.
- `ONTOLOGY_INSTANCE` nodes correspond to each HIV risk work in each HIV risk bucket.

Each of the aforementioned nodes were connected to each other via one of the nine different types of edges listed below.

- `TWEETED` edges from a `USER` node to a `TWEET` node, indicate the *author* relationship of the tweet.
- `IS_REPLY_FOR` edges from a `TWEET` node to another `TWEET` node indicate that a tweet is a *reply* for another one.
- `RETWEET_OF` edges from a `TWEET` node to another `TWEET` node indicate that the tweet is a *retweet* of another tweet.
- `FOLLOWS` edges from a `USER` node to a `USER` node indicate what other users the current user is *following* on Twitter.
- `MENTIONED_IN` edges from a `USER` node to a `TWEET` node indicate when a user is mentioned (a reference to his @ handle) in another tweet.
- `HAS_HASHTAG` edges from a `TWEET` node to a `HASHTAG node` indicate that the specific tweet contains the listed hashtag (#).
- `HAS_URL` edges from a `TWEET` node to a `URL` node indicate if and what URL is included in a Tweet.
- `HAS_RISK_WORD` edges from a `TWEET` node to an `ONTOLOGY_INSTANCE` node indicate the risk word the specific Tweet has been assigned to (can be multiple).
- `INSTANCE_OF` edges from an `ONTOLOGY_INSTANCE` node to an `ONTOLOGY_BUCKET` node indicate the bucket every risk word belongs to.

In order to facilitate analysis of the collected data at the network level we stored all the data as a graph in the Neo4J[3] graph database. Previously, without focusing on the relationships and interactions among the HIV at-risk users, we were able to see HIV transmission

[3] http://neo4j.com/

risk behaviors at the granularity of a geographical area and were able to perform longitudinal studies on the behavior of the concerned users. With a graph representation, we can derive more insights by exploiting the interactions and relationships among HIV at-risk users. For instance, once we conclude based on available data that user X and Y were indeed connected (e.g. due to HIV transmission), we can infer different patterns such as similarity in topics both X and Y tweet about, frequency and length of conversations between X and Y, and connectivity. If the two users are connected, then information such as the nature of the edges connecting X to Y, the existence of common hubs, etc. can easily be studied. Representing a social network in the form of a graph provides therefore a significant added value.

Neo4j, the graph database that we decided to use, is implemented in Java and provides numerous libraries and plugins to execute well-known graph algorithms on our data. Data stored in Neo4J can be queried using a query language called CYPHER.[4] For instance, to extract the top five most mentioned users in the HIV at-risk network, we can use the following CYPHER query:

```
MATCH p=((u:USER)-[r:MENTIONED_IN]->(t))
where not (t)-[:‘IS_REPLY_FOR‘]->(:‘TWEET‘)
RETURN u.name,count(p) AS num_mentions
ORDER BY num_mentions DESC limit 5;
```

Similarly, using CYPHER, it is possible to extract interaction chains between users who exhibit drug-risk behaviors and sex-risk behaviors on Twitter. Consider the following query:

```
MATCH p1=((n:ONTOLOGY_BUCKET)-[r]
-(m:ONTOLOGY_INSTANCE)-[r1]-
(t:TWEET)<-[r2:IS_REPLY_FOR*2..]-(t1:TWEET)
-[r3]-
(o:ONTOLOGY_INSTANCE)-[r4]
-(p:ONTOLOGY_BUCKET {id:’DrugBucket’}))
where n.id
in ["HomosexualTermsBucket","STIBucket",
"SexBucket","SexVenues"]
and not (t)-[:‘IS_REPLY_FOR‘]->(:‘TWEET‘)
RETURN count(DISTINCT t);
```

In our future work (see below) we want to further analyze HIV-positive users that are known in the network, and we believe that graph queries like the ones mentioned above could help find social circuits that are at risk of infection.

6. UNDERSTANDING RISK NETWORKS

Once we obtained the HIV at-risk social network graph, a subgraph of the entire Twitter graph, we wanted to understand if the network properties of this sub-graph match those of the actual Twitter social network graph. In order to investigate this rather open space, we decided to employ Exploratory Data Analysis (EDA) [26] and a series of specific questions that we run through our Neo4J infrastructure. To investigate peculiar characteristics of the HIV transmission risk network we compared the results of both the HIV at-risk network and the whole Twitter social network. In the remainder of this section we detail the results of our explorations and investigations.

6.1 Users’ tweeting trends

After filtering the collected tweets as outlined above, as of August 2015 we identified a total of 13,000 HIV at-risk users. While in our future work it will be important to continue analysis of their networks to better assess the degree of risk, we initially checked the the number of HIV risk tweets per user. Generally, any kind of social network shows a power law degree distribution due to the preferential attachment exhibited in such networks [5] and it is interesting to see a similar pattern in the HIV at-risk sub-network as well (Fig. 6). This shows that the extracted HIV at-risk sub-network is preserving some of the network property of power-law distribution of tweets.

We additionally investigated if the length of each tweet revealed anything interesting about the HIV riskiness of the tweet. Again we could not find any marked difference between the distributions.

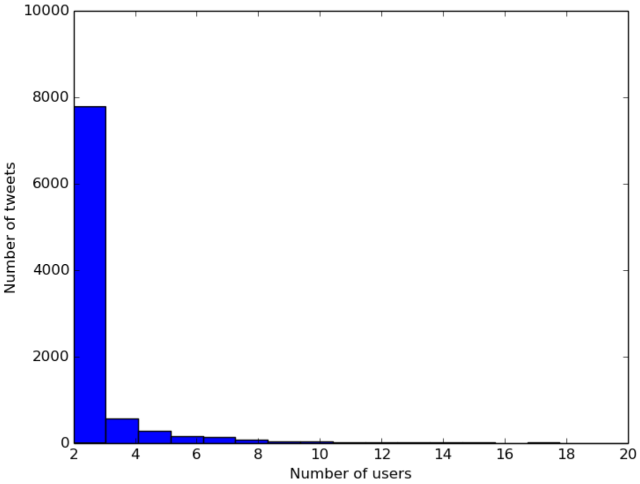

Figure 6: Distribution of HIV risk tweets across users

6.2 Tweeting and HIV risk factors

Given that our data collection was based on five different risk-factor buckets, we decided to see how these bucket are reflected in the collected tweets. We saw a major fraction of HIV transmission risk tweets falling within the homosexuality and drug categories (Fig. 7).

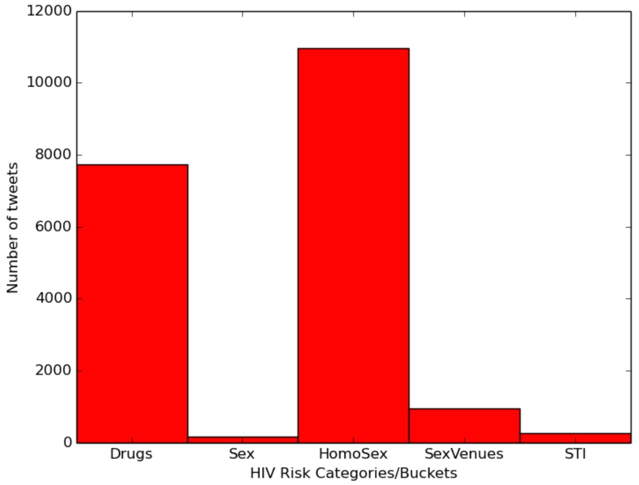

Figure 7: Distribution of tweets based on risk categories

[4] http://neo4j.com/developer/cypher-query-language/

After understanding that among HIV risk categories, the most pronounced were drug and homosexuality related behaviors, it was important to explore which of these categories could have the highest likelihood of occurring together. We therefore ran a comparative analysis across all buckets, and as illustrated by the confusion matrix in Fig. 8 we found that the risk behavior that were more often mentioned together were (*Homosexuality* and *Drug*), (*Sex* and *Drug*), and (*Sex Venues* and *Homosexuality*).

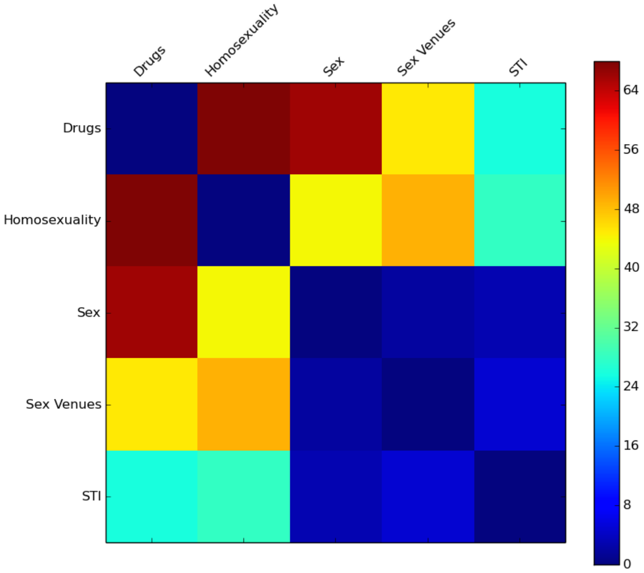

Figure 8: Co-occurrence patterns of HIV risk categories. The color bar (right) shows the color corresponding to the proportion of tweets in which terms from two risk buckets co-occur.

7. DISCUSSION AND FUTURE WORK

In the previous section we showed how we could exploit our computational infrastructure to query the data we collected from Twitter and associate this with HIV at-risk behavior. Our EDA approach showed how we can begin to better understand the interplay between various factors that make a tweet a HIV transmission risk tweet.

While we demonstrated the potential for this understanding, much remains to be done to truly characterize the HIV at-risk network.

Firstly, from the results shown in the previous section, we have affirmed that the HIV at-risk social network has similar basic properties as the underlying Twitter social network. The next step is reconstructing the real-world HIV transmission network, as previously done by our group [16], and associate it to the Twitter social network. To investigate this, we are currently conducting a comparative study with real-world networks wherein we request HIV-positive (i.e., infected) and HIV-negative (i.e., uninfected and status unknown) users to provide us their Twitter handles. The tweets from these Twitter accounts are used as the gold standard to understand how real-world HIV infected and uninfected users would actually behave on a social network. By augmenting the original HIV at-risk social network with these new Twitter accounts, and by superimposing the actual HIV transmission network on top of this augmented graph, we might be able to better understand how a real-world putative HIV transmission edge would translate into relationships and interactions in the Twitter social network. This will enable us to build a computational model that could run independently on all the user nodes in the Twitter social network graph to understand each user's susceptibility to acquiring HIV. This in turn will further refine the HIV heat-map which we showed earlier in Fig. 4.

Secondly, the HIV at-risk social network graph, which is entirely based on the HIV risk tweets, and is classified based on risk behaviors exhibited via the tweet content, might still show a considerable number of false positives due to the subtle differences in the intended meaning of the tweet. For instance, a tweet mentioning consumption of meth (a drug) may not always mean that the tweeter exhibited HIV risk behavior. To identify such tweets and to learn the underlying patterns in such false positives, we are introducing machine learning techniques, similar to the work done in [2, 4], well-suited for text classification using features including the content of the tweets, the description of the author profiles, the time of the day when each tweet was posted, the day of the week and the presence of negative terms like "not", "never" and so on.

8. CONCLUSION

We demonstrated in this paper the capability of Twitter to help public health and prevention efforts by acting as a radar for infectious diseases such as HIV. These kind of radars could act at a geographic level, using the geo-tagging information embedded in the tweets, or at the individual level using graph algorithms as the ones described above. In the process of building our Twitter radar, we have also defined a structured approach for collecting, cleaning, classifying, modeling, processing and deriving valuable intuitions from social networks for the purposes of digital epidemiology. As part of this paper we shared best practices in collecting and managing streaming data. We realized that processes like classification and cleaning of data are iterative processes and should be designed as such when building the data pipeline for managing social data. Finally, our exploratory data analysis revealed latent aspects of the social network that provide insights on the factors influencing HIV risk behavior.

All in all we see our work as a first step towards uncovering the incredible potential of linking social media with HIV risk behavior. We believe that the use of Twitter data as a real-time tool to characterize and monitor infections, and in particular MSM social network connectivity in San Diego is real. The research we presented here will lead to characterizing the relationship between users who tweet about high-risk behavior and derive their social and sexual networks. Our work the potential to produce major impact on a broad set of medical and behavioral research, opening up a new exciting wave of research in a field that is just starting to be explored as a support for public health. The outcome of this research will not only help us to characterize HIV at-risk networks, but will also act as a springboard for a modern methodology that exploits social media to better understand the spread of real-world diseases. If successful, we already foresee how the methodologies presented here could help develop other combined social media and epidemiological studies focusing on different diseases, in turn increasing speed and efficiency for tracking and containing other epidemics.

9. ACKNOWLEDGMENTS

This research has been funded in parts by the UC San Diego Center for AIDS Research (CFAR) NIAID P30 AI36214 (PI Weibel), the UC San Diego Frontiers of Innovation Scholars Program (PI Weibel), and the NIAID Award R24 AI106039 (PI Little). We would like to thank the AVRC team at UCSD and the local CFAR for giving us access to their resources and expertise in HIV/AIDS.

10. REFERENCES

[1] Benjamin M Althouse, Yih Yng Ng, and Derek AT Cummings. Prediction of Dengue Incidence Using Search Query Surveillance. *PLoS neglected tropical diseases*, 5(8):e1258, 2011.

[2] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches the Flu: Detecting Influenza Epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[3] John W Ayers, Kurt M Ribisl, and John S Brownstein. Tracking the Rise in Popularity of Electronic Nicotine Delivery Systems (Electronic Cigarettes) Using search Query Surveillance. *American journal of preventive medicine*, 40(4):448–453, 2011.

[4] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards Large-scale Twitter Mining for Drug-related Adverse Events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM, 2012.

[5] David R Bild, Yue Liu, Robert P Dick, Z Morley Mao, and Dan S Wallach. Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *arXiv preprint arXiv:1402.2671*, 2014.

[6] Logan Broeckaert and Margaret Haworth-Brockman. "You may have come into contact with...": HIV Contact Tracing in Canada. *Prevention*, 2014.

[7] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine*, 360(21):2153–2157, 2009.

[8] CDC. Estimated HIV Incidence in the United States 2007-2010. *HIV Surveillance Supplemental Report 2012*, 17(4), 2012.

[9] Hyunyoung Choi and Hal Varian. Predicting the Present With Google Trends. *Economic Record*, 88(s1):2–9, 2012.

[10] Nielson Company. State of the Media - The Social Media Report 2012. *nielson.com*, 2012.

[11] Alex J Elliot, Angie Bone, Roger Morbey, Helen E Hughes, Sally Harcourt, Sue Smith, Paul Loveridge, Helen K Green, Richard Pebody, Nick Andrews, et al. Using Real-time Syndromic Surveillance to Assess the Health Impact of the 2013 Heatwave in England. *Environmental research*, 135:31–36, 2014.

[12] Gunther Eysenbach. Infodemiology: Tracking Flu-related searches on the Web for Syndromic Surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, page 244. American Medical Informatics Association, 2006.

[13] James H Fowler, Nicholas A Christakis, et al. Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. *Bmj*, 337:a2338, 2008.

[14] Kelly J Henning. What is Syndromic Surveillance? *Morbidity and Mortality Weekly Report*, pages 7–11, 2004.

[15] Bernhard Hofmann-Wellenhof, Herbert Lichtenegger, and James Collins. *Global Positioning System: Theory and Practice*. Springer Science & Business Media, 2012.

[16] Susan J Little, Sergei L Kosakovsky Pond, Christy M Anderson, Jason A Young, Joel O Wertheim, Sanjay R Mehta, Susanne May, and Davey M Smith. Using HIV Networks to Inform Real Time Prevention Interventions. *PloS one*, 9(6):e98443, 2014.

[17] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *arXiv preprint arXiv:1306.5204*, 2013.

[18] Elizabeth L Murnane and Scott Counts. Unraveling Abstinence and Relapse: Smoking Cessation Reflected in Social Media. In *Proc. CHI 2014*, pages 1345–1354. ACM, 2014.

[19] Christian Napoli, Flavia Riccardo, Silvia Declich, Maria Grazia Dente, Maria Grazia Pompa, Caterina Rizzo, Maria Cristina Rota, and Antonino Bella. An Early Warning System Based on Syndromic Surveillance to Detect Potential Health Emergencies among Migrants: Results of a Two-Year Experience in Italy. *International journal of environmental research and public health*, 11(8):8529–8541, 2014.

[20] Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. Inferring the Origin Locations of Tweets With Quantitative Confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536. ACM, 2014.

[21] Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. You Are What You Tweet: Personality Expression and Eerception on Twitter. *J Res Pers*, 46(6):710–718, 2012.

[22] Zhaochun Ren, Maria-Hendrike Peetz, Shangsong Liang, Willemijn van Dolen, and Maarten de Rijke. Hierarchical Multi-label Classification of Social Text Streams. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 213–222. ACM, 2014.

[23] Marcel Salathe, Linus Bengtsson, Todd J Bodnar, Devon D Brewer, John S Brownstein, Caroline Buckee, Ellsworth M Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L Mabry, et al. Digital Epidemiology. *PLoS computational biology*, 8(7):e1002616, 2012.

[24] Gwendolyn Seidman. Self-presentation and Belonging on Facebook: How to Influence Social Media Use and Motivations. *Personality and Individual Differences*, 54(3):402–407, 2013.

[25] We Are Social. Digital, Social and Mobile in 2015. *http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/.*

[26] John W Tukey. Exploratory Data Analysis. *Reading, Ma*, 231:32, 1977.

[27] Sean D Young, Caitlin Rivers, and Bryan Lewis. Methods of Using Real-time Social Media Technologies for Detection and Remote Monitoring of HIV Outcomes. *Preventive medicine*, 63:112–115, 2014.