HIV Risk on Twitter: The Ethical Dimension of Social Media Evidence-based Prevention for Vulnerable Populations

Nadir Weibel, Purvi Desai, Lawrence Saul, Amarnath Gupta, Susan Little
University of California, San Diego,
9500 Gillman Dr., La Jolla, CA 92093, USA
{*weibel, pdesai, lsaul, a1gupta, little*}*@ucsd.edu*

***Abstract*—As of 2016 the HIV/AIDS epidemics is still a key public health problem. Recent reports showed that alarmingly high numbers of people in vulnerable populations are not reached by preventative efforts. Despite technology improvement, we are not yet able to identify populations that are most susceptible to HIV infections. In order to enable evidence-based prevention, we are studying new methods to identify HIV at-risk populations, exploiting Twitter posts as possible indicators of HIV risk. Our research on social network analysis and machine learning outlined the feasibility of using tweets as monitoring tool for HIV-related risk at the demographic, geographical, and social network level. However, this approach highlights ethical dilemmas in three different areas: data collection and analysis, risk inference through imperfect probabilistic approaches, and data-driven prevention. We contribute a description, analysis and discussion of ethics based on our 2-year experience with clinicians, IRBs, and local HIV communities in San Diego, California.**

***Keywords*- Twitter, HIV/AIDS, Social Media, Digital Epidemiology, Ethics**

I. Introduction and Related Work

HIV remains a significant public health problem as indicated by the numbers of new HIV infections, particularly in vulnerable populations such as Men who have Sex with Men (MSM). In the USA, recent reports show a general increase of 8% between 2001—2012 and 22% increase among young (13–24 yr old) gay and bisexual men, which in turn represent 72% of all new infections in that age group [1]. Similarly, the incidence of HIV in MSM in the United Kingdom has increased between 1990 and 2010 with an estimated mean incidence of 0.30/100 person-years between 1990–1997 and 0.45/100 person years between 1998–2010 [2]. The generalized approach to HIV prevention has focused on universal HIV screening [3], though the proportion of adults in the U.S. who have *"ever tested"* remains below 50%. In general, efforts to focus prevention messages and resources to those at greatest risk of acquiring infection are lacking.

On the other side, mapping the social and sexual network dynamics of MSM at greatest risk of acquiring HIV, and associating these networks with available, well-characterized sexual network characteristics of HIV-infected transmission networks may provide opportunities to evaluate novel targeted approaches to HIV prevention interventions. Specifically, use of molecular epidemiological methods have greatly increased our understanding of HIV transmission dynamics. By making use of the HIV sequence data derived from routine (standard of care) HIV drug resistance testing, it is possible to infer a partial transmission network. For instance, these data have been used to extensively characterize the San Diego Primary Infection Cohort (SD PIC) HIV transmission network and identify the network features of persons at greatest risk for HIV transmission within their first year following incident infection [4]. Simulations of these data also demonstrate that targeted antiretroviral therapy (ART) to those with the highest overall risk of transmission (based on an objectively derived *"transmission network score"*) resulted in a significantly greater reduction in HIV network transmission as compared to random ART [4]. However, while early and universal ART in HIV infected persons is the cornerstone of effective HIV prevention, these methods do not address the HIV acquisition risk among HIV-uninfected persons in the same sexual network.

Just as HIV transmission networks are used to identify high-yield prevention targets for ART interventions, we believe that social network data derived from social media may be used to characterize the at-risk social network structure, and may help to identify those at greatest risk for acquiring HIV infection. Among current social media platforms, Twitter – a highly trafficked social media network built on tweets and followers used around the world – is particularly interesting. It is used not only to connect with friends/acquaintances online but also to follow real-time events (such as the Arab Spring [5], or the extent of an earthquake [6]) and other information users find interesting. The potential to use Twitter as real-time HIV monitoring method was explored by Young and colleagues at UCLA who showed a relationship of tweets containing drug and sex related words, with HIV prevalence data reported by the US Center for Diseases Control (CDC) [7], [8]. These data demonstrated a significant correlation between higher numbers of HIV-risk communications and higher HIV prevalence within the county. This suggests that HIV-risk behavior that is shared by Twitter users online via tweets to their followers may be used to infer regional network characteristics of HIV risk.

This initial work shows the incredible potential of linking social media with HIV-risk behavior and potentially use Twitter data as a real-time tool to characterize and monitor HIV-at risk social networks. Particularly compelling is the

URI: http://hdl.handle.net/10125/41370
ISBN: 978-0-9981331-0-2

opportunity that publicly available tweets offer, as a way to unobtrusively collect and analyze this data, identify potential at-risk individuals and isolate their social networks, and then intervene with appropriate preventative efforts, either at the individual or at the community level. This approach presents the opportunity to open up avenues on one of two fronts: (i) linkage to care and free HIV testing advertisements, and (ii) identification of new groups (geographic, socio-demographic, etc.) that can be reached to implement targeted prevention campaigns. However, what initially seems like a moderately complex *big data* research program, with clear applicability into the real world, turns out to entice a number of critical ethical questions, in particular in relationship with the vulnerable HIV/AIDS populations, possible issues with stigma, as well as general Digital and Social Media (DSM) analysis and inferences.

The contribution of this paper is to introduce the ethical issues that emerged from the development and deployment of our PIRC-Net infrastructure [9] aimed to infer HIV risk from publicly available tweets, and then critically discuss those issues in terms of three specific point of views: data collection and analysis, modeling and risk inference, and preventative efforts. In the remainder of this paper we will first introduce our PIRC-Net infrastructure, including our data collection and analysis strategies. We will then highlight opportunities for prevention efforts based on the data that our research is generating. Finally we will move into a critical discussion of the ethical dimension of our research across the three areas outlined above.

II. The PIRC-Net Infrastructure

HIV's primary route of transmission is governed by connections within social or sexual networks, therefore analyzing such networks can inform public health measures to contain the spread of HIV. However, these networks tend to have a dynamic structure, typically becoming apparent only after the long incubation period between HIV transmission and disease state. Moreover, due to HIV's low transmission rate per contact, infections typically only involve subsections of a social network. Traditional methods of defining network features through interviews and partner-tracing are not proving as effective [4]. On the other side, given how people's social media footprint mirrors their real life to a large extent [10], [11], real-time analysis of social networks could help build an infection surveillance radar of HIV transmission risk behavior. Twitter data have the potential to equip us with location and population based HIV risk behavior indicators to infer transmission networks which could help provide an early warning indicator for HIV risk [12]. Research leading to characterize the relationship between users who tweet about high-risk behavior and derive social and sexual networks has the potential to produce major impact on a broad set of medical and behavioral research, and open up a new exciting wave of possibilities.

In order to exploit this exciting space, we built PIRC-Net a computational infrastructure aimed at collecting data from Twitter, filter this data for possible risk of HIV, and then identify relevant features at the social network level to enable better characterization of at-risk networks [9]. PIRC-Net's goal is to enable initial exploratory analysis of the emerging social structures to unravel the relationships between different HIV risk behaviors. Various dimensions of this data such as geography, demographics, and social groups have been used in PIRC-Net to model HIV risk.

A. Data Collection

Twitter provides an Application Programming Interface[1] (API) to programmatically access users' data. Along with the standard APIs, Twitter provides a Streaming API that creates a long-standing connection between the client and the server, and streams the incoming tweets to the subscribing clients. In order to capture tweets from San Diego alone, we used Twitter's *Filter Hose* API which allowed us to collect geo-tagged tweets that are generated within our San Diego geocoded bounding box in real-time.

The tweets collected using the Streaming API are continuously pushed onto a MongoDB[2] database that stores them and enables easy access for analysis. MongoDB was used mainly owing to the ease of storing semi-structured documents, high write throughput, and support for native map-reduce queries for performing on-demand aggregations.

To identify HIV tweets that indicate risk behavior, all tweets from San Diego County are filtered to create a smaller corpus of tweets based on the presence of certain HIV transmission *"risk words"* in their tweet content. A risk word is essentially a term considered to be positively correlated with HIV risk behavior. A small group of clinical collaborators –faculty and trainees working in the field– with daily experience in the setting of prevention and treatment, as well as outreach in the community acted as domain experts and helped us to define five broad categories of HIV risk words. These *buckets* were populated with words frequently used in the local San Diego community.

1) **Substance Abuse Bucket**
   E.g. meth, ice, snow, cocaine, party&play
2) **Sex Bucket**
   E.g. creampie, cottaging, bronco, party&play
3) **Sex Venues Bucket**
   E.g. loft, redwing, bourbon street
4) **MSM Sexual Behaviors Bucket**
   E.g. homo, gay, queen
5) **Sexually Transmitted Infection Bucket**
   E.g. syfy, drip, gleet

Data collection is followed by data filtering to eliminate false positives, and then pulling from Twitter all related users

[1] http://dev.twitter.com
[2] http://www.mongodb.com

who either re-tweeted a tweet or were mentioned in one. These data are stored on a separate MongoDB collection. In this initial filtering phase, it was important to understand the context in which the keywords are used to determine the at-risk nature of the tweet. For example, a tweet such as "I love ice coffee" could be miscategorized due to the presence of the keyword `ice`. Our filtering algorithms discard such false positives to ensure better signal-to-noise ratio. We discuss this in detail in the next subsection.

B. Data Cleaning

The data collection process filters the data based on the presence of keywords from the five HIV risk categories. However, on a microblogging social network like Twitter, the text is limited to 140 characters. This leads to the use of a lot of text shortening, abbreviation, and emoticons. Additionally, even though our data collection process filtered the incoming tweets based on the presence of certain risk keywords, these risk words could mean something entirely different in different contexts, and hence at times be misfiltered as at-risk. In order to filter out such false positives, we defined *inclusion* and *exclusion* lists based on the co-occurrence of words with these risk keywords. For instance, the exclusion list for the keyword *"crack"* (slang for meth/drug) would include *crack me up, crack myself up, crack up, crack open, crack of dawn*. A tweet with a phrase that reads *"crack me up"* would pass our first crude filter for tweets tagged with risk words based merely on the presence of certain predefined keywords. However, the second level filter based on inclusion/exclusion lists would discard such false positives based on the co-occurrence of words. The inclusion list on the other hand would include words that if co-occurred with the keywords under consideration would allow the tweets to pass through this second level of filtering.

This data cleaning phase led to the reduction of tweets that passed the initial keyword filter based by 60%, thus improving the signal in the data. While the two-stage filtering process significantly brings down the number of irrelevant tweets that are considered at risk, a fairly large number of false positives still remain in the corpus. To understand the veracity of the filtering process, a domain expert on our team manually assigned *"Positive Risk"* and *"Negative Risk"* labels to a subset of our corpus of already filtered Tweets. We observed that the final manually labeled data comprised of 27% true positives (positive risk) and 72% of false positives (negative risk) with respect to our filtering approach, indicating still a very high number of wrongly classified tweets, mostly due to tweets that did not follow our inclusion/exclusion criteria list.

C. Machine Learning Filtering

This initial analysis was performed on a randomly selected small corpus of our data, and we assumed that the rest of the data would show similar behavior in the context of false vs. true positives. Given the large size of the corpus we decided to use a Machine Learning (ML) approach to build a better filter. We recruited 30 domain experts from our local HIV at risk community who manually labeled a larger amount of tweets on a 4-point scale (high-risk, risk, low-risk, no-risk) or an additional *"I do not understand"*. As expected, the majority of our tweets was classified as not showing any risk behavior. We used this data to train a classifier using a variety of algorithms, and then test the results against the labeled data. While it is outside of the scope of this paper to discuss our ML approach in detail, applying a Multinomial Bayes approach [13] that looked at consecutive word (unigrams, bigrams and trigrams) in combination with our risk words, brought the true-positive/false-positive ratio from 3/8 to 17/1, resulting in a 45 times better accuracy.

D. PIRC-Net Social Network Analysis

In addition to the raw analysis of the textual at-risk features of tweets, we decided to capture user connections within the social network to provide additional signal for HIV risk analysis. The user-tweet and user-user connections were modeled in a graph in the form of nodes and relationships with the help of Twitter's APIs accessing publicly available data. In order to store this network based data we built a Twitter network model and used the Neo4J database management system.[3] Our network model was based on seven different concepts:

1) **USER**, Twitter users at HIV risk based on their tweets.
2) **TWEET**, HIV tweet containing risk words in Twitter.
3) **HASHTAG**, Hashtag used in the tweet tagged with HIV risk words.
4) **URL**, URL being referred to in the tweet containing risk words.
5) **FOLLOWER_USER**, Set of users that follow each of the at risk users (So FOLLOWER_USER may or may not be an HIV risk Twitter user).
6) **ONTOLOGY_BUCKET**, The five risk buckets we defined above.
7) **ONTOLOGY_INSTANCE**, HIV risk word in each HIV risk bucket.

Each of these concepts are considered nodes in our network model and are connected via edges to each other. We defined nine different types of edges:

1) **TWEETED**, USER to TWEET: author of the tweet.
2) **IS_REPLY_FOR**, TWEET to TWEET: reply to another tweet.
3) **RETWEET_OF**, TWEET to TWEET: retweet of another tweet.
4) **FOLLOWS**, USER to USER: following on Twitter.
5) **MENTIONED_IN**, USER to TWEET: user mentioned in tweet (with reference to @ handle)

[3] http://neo4j.com/

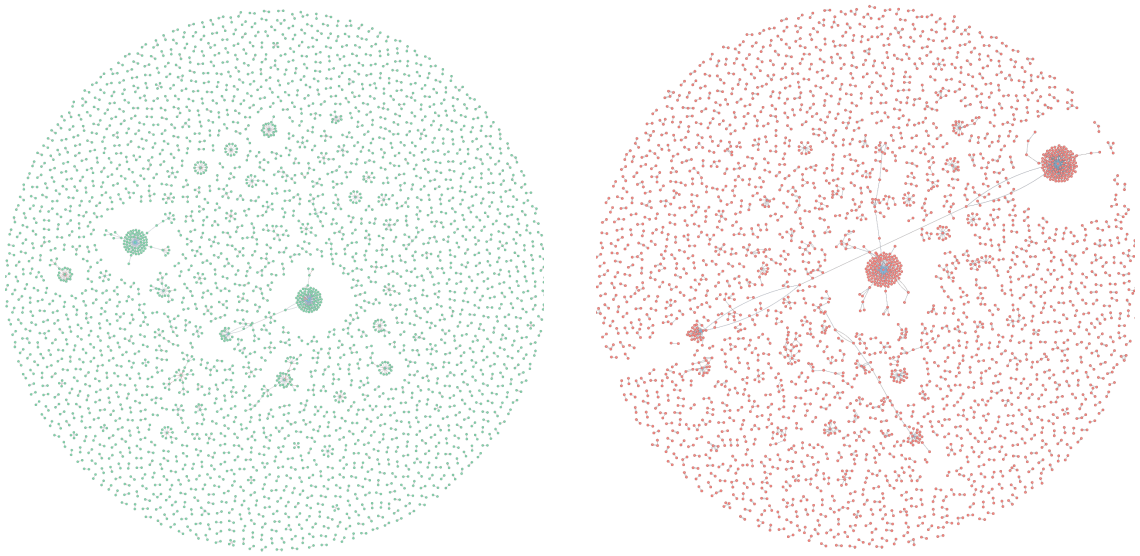

Figure 1. Social network and connectivity visualization of at-risk users based on conversations containing risk words (extracted from Neo4J). Left: overall risk-tagged conversations. Right: conversations tagged with terms suggesting male-male sexual behaviors (MSM). Clusters of users who engage in risk-tagged conversations clearly appear in both graphs. The MSM-tagged conversation also shows interesting cross-clusters connections.

6) **HAS_HASHTAG**, TWEET to HASHTAG: tweet contains listed hashtag (#).
7) **HAS_URL**, TWEET to URL: URL included in a tweet.
8) **HAS_RISK_WORD**, TWEET to ONTOLOGY_INSTANCE: risk word assigned to tweet (can be multiple).
9) **INSTANCE_OF**, ONTOLOGY_INSTANCE to ONTOLOGY_BUCKET: bucket the risk word belongs to.

We used this network-based information along with the text based information to help analyze HIV risk from both a single user and a network or community perspective. For instance, one could identify *"hubs"* or *"influencers"* from the network, based on the several different types of social relationships described above, or similarity in HIV riskiness between two users who share a connection. Graphical representation of the risk-network in terms of social network relationships could therefore offer significant added value. We will expand on the possibilities for prevention and interventions that these data yield in the next section.

III. Data-driven Prevention Opportunities

One of the key motivators of our research is answering the question *"Can Twitter's network data be used as a tool to infer the social network of individuals that are at high risk of acquisition or transmission of HIV?"* and then intervene in the real world to try and decrease the risk of these individuals and groups. Our computational infrastructure and network model allow us to explore types of social relationships on Twitter that could possibly be good indicators of real life user-user connections. In particular, we decided to explore the following four types of social relationships derived from our HIV risk network that seem to indicate strong user-user connections. In turn, one or more of these relationships could potentially be good indicators of real world contact and therefore real risk and enable us to extract risk-networks in a more reliable way:

1) **Conversations**: Connections between users that engaged in HIV at-risk conversations with one another via exchanging tweets. Every pair of users that engages in a direct exchange of tweets tagged with risk words is said to be part of *"risk-tagged conversations"*. We can derive additional attributes for each conversation, such as frequency, average length, risk category, type of conversation chain, etc.
2) **Geographic + Temporal Co-location**: Connections between users that were co-present in a spatial+temporal sense. Every pair of users that were within 1 mile apart from one another in a time window of 1 hour are said to have been geographically and temporally co-present.
3) **Mentions**: Connections between Twitter users who *"mention"* (using the *"@"* sign) another individual in a tweet as a mechanism to directly address that tweet to the mentioned individual.
4) **Follows**: Connections between users who *"follow"* other individuals as a mechanism for the followers to consume content generated by the followees.

While it is still unclear to what extent these four social relationships should be exploited for effective prevention efforts, it is possible to start exploring the risk network by looking at those social networks graphically. Figure 1 shows an example of an extract of conversations related to HIV-risk overall, and specifically in the setting of one of our risk category (conversations tagged with MSM sexual behavior).

Besides the overview visualization shown in Fig. 1, it is also possible to investigate those clusters in details and try to better understand the characteristics of those conversations. Figure 2 shows such an example, where different kinds of relationships can be seen, namely a hybrid substance abuse-MSM risk hub, a smaller substance abuse hub, a bidirectional, and two uni-directional conversation patterns.

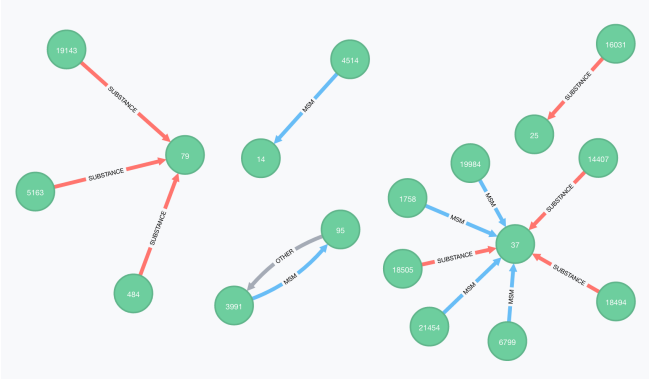

Figure 2. Zoomed View: Graph Visualization of Conversations within the HIV Risk Network. The green nodes indicate the user nodes identified by a system-generated identifier to protect the user's anonymity. The edges between these nodes indicated conversations, color-coded and labeled with the risk category that the conversation belongs to.

The advantage of our native network model, allows us to explore the data at finer levels of granularity. Besides the risk category, conversations networks can be easily extracted and visualized based on the frequency of the conversation between two users, on the average length of these conversations, and so on.

Another interesting way to look at the HIV tweets containing risk words, as well as their authors, is the geographical point of view. Every tweet in our continuously updating corpus is geo-tagged, meaning that we know exactly where it originated. This allows us to infer higher or lower presence of tweets with risk words on the territory. Figure 3 illustrates an example of how the density of tweets containing risk words can be overlayed on a map and can indicate potential areas that are more at risk than others. Interestingly, while some of these areas might be well-known to be at risk, others emerged as new potential, and unknown, risk areas.

The geographical information can also be exploited in combination with our network model to investigate co-location. Establishing co-location from an individual's social footprint can be of particular help to draw more direct connections from users social connections to their real world connections. These links can also help further characterize potential risk-venues that emerged from the geographic-only visualization. Figure 4 shows the graph visualization of user-user connections based on co-location derived using Neo4J. We can observe that several clusters are interspersed within this network and how the nature of this network differs from that observed in the conversations network.

Figure 3. Heatmap showing the geo-location of all tweets containing risk words in San Diego county from January 2016 to June 2016. Big clusters such as downtown San Diego, Hillcrest and Northpark (the big blob and smaller blob in the center) and the border region (at the bottom) are well known. Other spots in north and east county (e.g. Poway/Encinitas, Oceanside and Julian) emerge as new potential risk areas, not typically considered by preventative campaigns.

Finally, Twitter is also characterized by two additional more explicit networking information, *mentions* and *following*, which are both actively defined by users. Once again, our model and infrastructure allows us to explore these relationships as shown in Fig. 5 and 6.

Since users tend to follow those whom they either connect with in real life or whose opinions resonate with their own [14], [15], the follower/followees sub-network within the HIV risk network (Fig. 5) can provide useful cues towards understanding real-world user connections and prevention opportunities in terms of highly-connected networks.

Twitter also allows a user to directly address other users via tweets using their Twitter handles (e.g. *"@alice Let's party tonight!"* is a way to address this tweet directly to Alice). Social relationships based on user mentions such as the one shown in Fig. 6, could be another indicator of real world user-user connections since they indicate an alternative communication strategy across users.

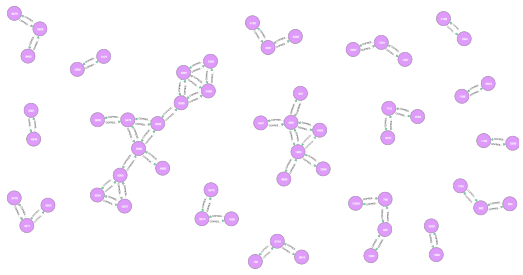

Figure 4. Co-location network visualization extract. Each purple node here represents a user, identified by a system generated integer to protect the anonymity of the users. The edges represent the *"colocation_with"* connections.

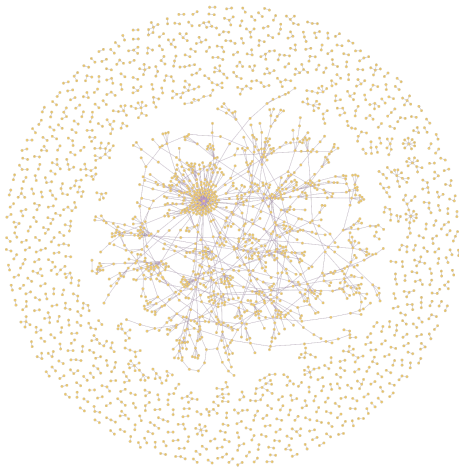

Figure 5. Following relationships in the HIV at-risk network. Every node in this graph visualization stands for a user in the HIV risk network. *"FOLLOWS"* directed edges join every pair of users where one of them follows the other. While the previous networks mostly showed disjoint clusters, this network highlights an interestingly unique form of clusters where most of the dense clusters are interconnected amongst themselves.

IV. Reflecting on Ethics for HIV risk on Twitter

Social network analysis and ethical dilemmas have been a vivid forum of discussions since the inception of the field of sociometric research [16] and then social network analysis [17]. Benefits and risks have been discussed by researchers and institutional review boards for years, and the impact of online social media was investigated already in the mid-1990s [18]. In discussing the ethics of social network analysis, benefits are typically abstracted from individuals to the level of society or humanity (i.e. *"how many lives can be saved"*) which is often used as the rationale for most of this research.

However, the relative balance of individual risk (related to loss of privacy) and public health benefit (related to use of network transmission data to guide public health strategies) remains a subject of discussion and debate among experts in the field. Already in 2005, Kadushin highlighted two

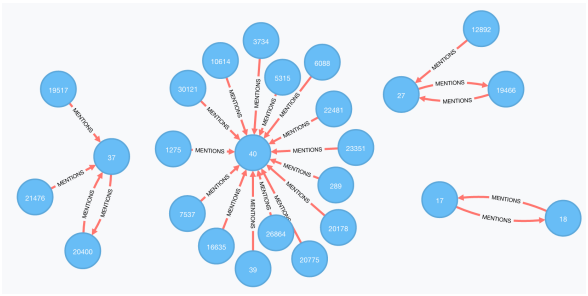

Figure 6. Mentions network as a subset of the HIV at-risk network. *"MENTIONS"* edges join every pair of HIV at-risk users where one of them mentions the other. Each of these edges is also characterized by a mentioned 'frequency' property. Edges are directed, the source node representing the user mentioning the target user.

important ethical issues that arise when investigating social relationships [19], i.e. the level of accuracy that should be provided so that reliable results can be presented (either at the individual level or as a community), and considerations around the results of social network analysis being subject to judicial review. Hoser and Nitschke additionally discussed how privacy issues and the increasing awareness on privacy, are critical in a connected world, but also how this is difficult when users themselves *"give away"* their data for free [20]. Eysenbach and Till stressed how the ethical issues are different if data is made available within a private vs. a public sphere, especially when connected to clinical or biomedical research [21].

Nowadays, after 10–15 years, we still face the same issues, but with massive amount of data and more capabilities to analyze these data and make inferences about them. Social media platforms make it easier to collect these data, and the existing APIs allow researchers (like us) to start exploiting these incredible data. In this section we would like to reconsider the dilemmas that researchers investigated in previous years, specifically highlighting and discussing the ethical dimension arising from our Twitter-based research on PIRC-Net. As mentioned above, the development of our platform, the data collection strategies, the storage of the collected data, our analyses, and finally our inferences onto the HIV/AIDS vulnerable populations all raised critical ethical concerns that need to be taken into serious consideration and need to be evaluated against the benefits of this research in terms of improved preventative efforts based on data-driven approaches. We do so here by analyzing our research in a structured way, looking at the ethics of data collection and analysis first, then discussing model-based inferences, and finally exploring ethics in prevention strategies.

A. Ethics of Data Collection and Analyses on Twitter

The initial effort in our research, and also the first step of our computational platform, focused on collecting geo-tagged data from Twitter through the available Twitter APIs. This step was then followed by our filtering process aimed at limiting our analyses to a corpus of tweets containing HIV risk words with a low probability of false positives.

1) Retrieving and Storing Publicly Available Data: The first ethical question that we encountered was directly related to the collection and storage of the data from Twitter. Given that we are only accessing publicly available Tweets that the users shared online and made accessible to everyone, our assumption was that the retrieval and local storage of this data posed no particular risks. However, a short exchange with our university's human subject protection office (or IRB) highlighted several potential risks. The IRB stated how it was necessary to *"Provide a more thorough analysis of the potential risks"* and how *"additional risks to users such as reputational and social risks should be outlined, as very sensitive material is being collected and analyzed"*.

While it is clear that we are working with a vulnerable population, it is not clear why the tweets shared by the users need to be considered very sensitive, and what *additional* reputational and social risk would be created in the process of collecting this data. As our work demonstrated, we were able to find many tweets exposing at-risk behavior (i.e. MSM sexual practices) that users choose to share online. Papacharissi and her colleagues remind us that people who decide to broadcast on a public forum like Twitter, did not necessarily think about their tweets being readily available for anybody to analyze and scrutinize [22]. Is this enough to warrant the enforcement of additional measures to collect and store the data in a *more* ethical manner, even if analysis of the data is conducted in an anonymized and aggregate way? We believe that this is heavily context-dependent and we need flexible guidelines on how to operate in this setting.

2) Anonymization and Privacy Procedures: During analysis of the tweets, depending on their content, we assigned a particular HIV risk value to single tweets and automatically also to their authors, making this information available to our research team. While anonymization of the author, as well as any other identifiable user information (such as mentions) is an important steps to warrant protection to the users, are other information of this *tagged* tweets to be masked?

IRB policy was to further anonymize the tweets and stripe them of any information that would allow *"quoted text to expose private information through public channels that can be searched for"*. This meant to transform the tweets by applying specific 'Name Entity Recognition' Natural Language Processing filters, such as the one implemented by MITRE,[4] to attempt to remove Personally Identifiable Information (PII). Additionally, in an effort to limit content reconstruction and linkage to the original users who posted the message, we removed stop words and punctuation from tweets, uniformed them to a lowercase form, and applied stem words reduction techniques. The resulting tweets were not linked to the original ones, but it was also not possible to understand their meaning anymore, making any further analysis on the tweets impossible.

3) Crowd-Sourcing and Expert-Sourcing: In order to obtain a labeled data-set, we initially planned to crowdsource the task on a mechanical turk platform[5] where turkers would manually assign labels to the tweets based on their perception of the tweet's HIV risk. Our ethical protocol mandated that the anonymization policy outlined above would have to be enforced in order to share the tweets through an online platform like mechanical turk. However, the aggressive de-identification of their content made turkers unable to understand the tweets and therefore to perform their labeling. Given that tweets are already in the public domain, we wonder to what extent this is necessary and what alternative could have been proposed? Perhaps framing the turker questions in a way that was not leading to inferences about HIV/AIDS would have been sufficient? In our case given the strict management of privacy/anonymization we instead recruited local domain expert to conduct a small scale labeling experiment on site, under our supervision. We judged that in this controlled situation aggressive anonymization policies were not necessary, so tweets were only removed of any direct identifier (author and mentions).

4) Judicial Review of Illegal Risk Behaviors: Given the inherent search for HIV risk behaviors, our research is predisposed to uncover behaviors that might also be illegal, especially in terms of substance abuse terms. Although this was not our goal and none of the analysis algorithms we developed tried to uncover illegal risk behaviors, tweets mentioning distribution of illicit drugs emerged in an anonymous way during our labeling and review activities. For instance a tweet described a neighbor's use of marijuana and methamphetapine in the courtyard, while another user, a possible drug dealer, tweeted to his followers to *direct message* him/her if they wanted to buy methamphetamine.

Ethical guidelines typically state that researchers have no legal guarantee of confidentiality in the matter of illegal behavior. Unless a certificate of confidentiality is obtained, if a researcher collects information about illegal activities, the researcher's data may typically be subpoenaed by the authorities. On the other side, researchers have no legal obligation to report these cases, especially if the presence of this data in their corpus might not be obvious. In addition, the data we are considering here are in the public domain, so what does subpoening these data really mean in this case?

Since these data are collected in big quantities, they are mostly analyzed by machine learning through an automatic computational infrastructure that anonymized the data as part of the process, so that behaviors that could have criminal justice consequences are not linked to the anonymous Twitter users. On the other side, the flagging of this information *could be* designed into the algorithms which could then extract and classify illegal behavior, similarly as we are classifying at-risk HIV behavior. We wonder to what extent this needs to be regulated and what is the most appropriate ethical framework to apply here? We believe that legal regulations have not been updated to reflect changing communication practices and available social media platforms, and must begin to take into account public available data.

B. Ethics of Model-driven inferences

Our work on PIRC-Net is based on data that is publicly made available through Twitter for reasons different than being classified for HIV at-risk behavior. Inherently this data is inaccurate and the inferences that we are making about the authors and their social networks are based on an artificial model that we built and tested on this data. Although this model might actually be effective,

[4]http://mist-deid.sourceforge.net

[5]http://www.mturk.com

we believe that we should always be cautious in the way we present potentially inaccurate results. Let us look at three assumptions that lead to three different inaccuracies.

1) Incomplete Data: Our strategies for collecting data are incomplete and inaccurate already in their design. In order to inform local intervention strategies, we only consider data from San Diego County. This means putting a filter on the Twitter API that will only deliver data that has been geo-tagged with a location within those boundaries. However, research has shown that only less than 1.6% of the tweets from Twitter sample hose are geotagged [23]. Additionally, while we filter for *verified* Twitter accounts to discern between organizations and individuals who are posting the tweets, this assumes that all organizations actually registered and verified their account with Twitter.

2) Text-based Analysis Strategies: The text-based filtering technique that is at the basis of our approach is grounded into an algorithm matching a set of key words manually defined to represent our five risk buckets, to the content of the single tweets. While this intuitively makes sense, the generation of this word list is a key component that can influence the resulting classification. In our case this has been defined by our clinical and domain experts, but it is hard to understand how universal these terms are. While we are planning to continuously updating the risk word list, the specific background and demographics of the domain experts contributing to populate this risk word list, as well as the different language slangs that are used in the local community critically influence the resulting filtered data.

3) Probabilistic Filters and Machine-Learning: Similarly as described above, our application of machine learning is based on manually labeled data. We demonstrated the effectiveness of our approach in reducing false positives, but this is still based on a relative small number of input tweets and a group of domain experts who might introduce specific cultural, language or demographic biases. The resulting filters are probabilistic and it is very hard to say how these results reflect reality in an accurate way.

Incomplete data, manually driven analyses, and probabilistic data filtering are approaches that only approximate the characterization of at-risk networks. While this represents an important advancement towards a data-driven approach, we must always keep these limitations in mind when we make inferences on specific users or communities. We believe that these methods could be useful to better understand and characterize networks of people at risk for HIV infection, however it is not clear what ethical framework needs to be used to frame these results in a way that does not over-interpret them.

C. Ethics of Data-driven interventions

This project unequivocally showed potential new pathways to drive interventions in the real world based on real-time collection of data on social media. While this is exciting, we need to reflect on how to further exploit this information, and which of the possible approaches presented earlier in Section III is ethically acceptable. In particular we would like the research community to reflect on the real linkage between social media and real world in the setting of HIV risk and prevention, on the meaning of geographically-based and network-based interventions, on the differences in addressing individuals vs. groups at risk of HIV infection, and on the opportunity to exploit emerging hubs and influencers in the network as vehicles for prevention.

1) Social Media vs. Real World: Our analysis of social networks and their potential for intervention is based on the assumption that we can reflect social media links to the real world. While this has been demonstrated on very specific behaviors in other settings [24], [25], it is unclear to what extent we can establish a clear relationship between online at-risk communication and real-world risk. If this relationship is not clear, how should we act on its inferences? Is it ethically appropriate to build an intervention around weak signals based on only approximate models? If we reflect on how prevention campaigns are developed right now, we realize that they are mostly based on partial views of the world and intuition. Is it acceptable to follow the same approach for social-media driven interventions?

2) Geo-based Interventions: The availability of geographical information that are stored with the tweet we collected through PIRC-Net, opens up opportunities to intervene in specific locations on the territory. What does this mean for the people living in those areas, especially when we uncover a new location? Is it acceptable to *tag* a location as being potentially at higher risk of HIV to increase the possibilities to reach individuals who happen to be in that region when tweeting about at-risk behavior? Can we intervene in specific venues (e.g. bars, restaurants), and what is the right way to do it?

3) Network-based Interventions: Similarly as for the geographical localization, our tools allow us to uncover new structures in the social fabric of the Twitter users who tweeted about HIV risk behavior. To what extent it is acceptable to reach out to those social networks, and what is the ethically correct way to do that? Should we collectively address high-risk networks as a whole? Should we identify single users and reach out to them? Moreover, how subtle should our message be? Is it ethically correct to make it evident that we have identified some behavioral patterns that are indicative of HIV risk in this specific network? Or should we use a higher-level, general strategy that although specifically targeted, might appear unrelated to the surrounding network?

4) Hubs and Influencers: From our initial explorations, it is apparent that the more data we collect, the more social structures appear in our PIRC-Net networks. Hubs and influencers (i.e. individuals central to the social network,

typically followed by many other users) are emerging as links for specific networks, or as a way to bridge between different networks. Intuitively we would like to exploit these important nodes within the network to spread our interventions. However, should we really try to do this? How can we approach these critical individuals or organizations in an ethically correct way, without pointing fingers or having a *big brother* effect? What is the approach that maximizes protection and respect of the participants on one side, and the chances of a successful intervention on the other side?

V. Discussion

The development and deployment of our PIRC-Net infrastructure, allowed us to identify and highlight ethical dilemmas of research to characterize HIV at-risk behaviors on Twitter. Admittedly, we have raised more questions than we provided answers, but this is intentional. We feel that we are still defining and experimenting with the correct ethical framework for social-media based big-data research, especially when vulnerable and complex populations such as HIV at risk individuals are at stake. Additionally, the application of machine learning approaches and social network analysis that is enabling researchers to create partial inferences on the real world behavior of individuals and groups, further complicates things. This approach might have different impact if used to characterize a population, or to drive public health intervention or campaigns, and we feel that more experiments and more data need to be collected on the ethical dilemmas around those different goals.

We see our work in this paper as a way to continue the discussion that is developing on defining ethical frameworks for social media research with vulnerable population. We hope to be able to continue these reflections and propose concrete guidance in the near future based on our experiences in social media research around HIV risk, especially by comparing and contrasting those experiences with the recent increasing number of attempts to define guidelines (ethical and beyond) for social media research. For instance we are looking at how to integrate our observations in the guidelines that Rivers and Lewis created for Twitter studies [26]. Those guidelines build on six different points:

1) Study designs using Twitter-derived data should be transparent and readily available to the public
2) The context in which a tweet is sent should be respected by researchers
3) All data that could be used to identify tweet authors, including geolocations, should be secured
4) No information from Twitter should be used to procure more data about tweet authors from other sources
5) Study designs that require data collection from a few individuals rather than aggregate analysis require Institutional Review Board (IRB) approval
6) Researchers should adhere to a users attempt to control his or her data by respecting privacy settings

Although this is certainly an interesting framework, it is hardly universal. Some of the guidelines make a lot of sense (i.e. #3 and #6), and are general enough to be applied widely. However, we feel that most of them represent a big hurdle for the research we presented here. For instance, 'making study designs available to the public' (#1) is hard if we are working on a collective risk radar. It is not feasible to publicly advertise the study in the hopes that those people that we are actually anonymously studying happened to read about it. Just making the general public aware of the research does not ensure targeting the right people. Furthermore, 'respecting the context of a tweet' (#2) would preclude making any inference about the content of the tweet (e.g. the HIV risk of the behavior represented by the tweet), as well as other possible questions on the collective behavior, making any real research that goes beyond simple textual analysis very hard. Moreover, 'not collecting additional data about the authors of a tweet' (#4) implies that we should not explore their social connections. Defining a priori that a study will require IRB approval if data collection is specific around few individuals rather than based on aggregate analysis (#5) can be challenging as researchers might discover only after having collected and analyzed large datasets that there are actually important aspects to consider related to specific individuals or groups. Based on the current regulations this would mean that the large datasets that researchers already collected would not be usable for these additional analyses.

More recently, Vitak and colleagues [27] reported on their analysis which surveyed 263 online data researchers, investigating their beliefs and practices around ethics and privacy in their own research. What their analysis made apparent is that while researchers in general try to aspire to the ethical principles that are at the basis of human subjects research, such as the Belmont Report [28], their practices are much more nuanced and do not fit easily in this static and rigorous framework. We believe that this is potentially due to the varying nature of the research methods, and questions that single research projects and specific investigators are exploring. The *ethics heuristics* approach that they proposed seems to scale better than other models, especially as a support of the current explorations, however our project shows that sometimes it is not clear how to effectively address some of these heuristics. For instance, Transparency is hard to address when it is not clear if and how one should reach out to specific subsets of the analyzed network

As an alternative, Hutton and Anderson [29] present an architecture for ethical and privacy-sensitive social network experiments. They propose to automatically implement ethics and privacy policies as part of their computational infrastructure. This seems like the right way to apply ethics at a system level; yet it is unclear how those ethics policies are defined and it is difficult to see how these should be implemented as part of a computational system. We believe that taking their approach further would enable many of

these computational systems, including PIRC-Net to exploit an integrated infrastructure. Potentially this approach could be merged with a more nuanced ethics policies as proposed by Vitak et al. to enable a more flexible and adaptable framework, which could suit the expanding landscape of social media studies.

VI. Conclusion

Developing effective ethical practices for big-data social-media research is hard. Despite the initial attempts to define unique frameworks, it seems that a *"one size fits all"* does not really work. This is particularly apparent when we approach the research from a different point of view such as informing prevention efforts for vulnerable population through analysis of social media interactions. In this paper we showed how the development and deployment of our PIRC-Net infrastructure allowed us to inform the practice of prevention based on HIV risk on Twitter, and how this opens up a number of interesting approaches to better serve the community and potentially help controlling the HIV/AIDS epidemics. Our focus has exposed the ethical dilemmas that these opportunities bring along, and how these complex socio-technical problems have no easy solution. We believe that our continuing experience in this setting will provide more guidance towards extending heuristic-centered approaches into more agile and easily tailored ways to protect human subjects, while preserving the ability of researchers to effectively support public health efforts.

References

[1] CDC, "Estimated HIV Incidence in the United States 2007-2010," *HIV Surveillance Supplemental Report 2012*, vol. 17, no. 4, 2012, http://www.cdc.gov/hiv/group/msm.

[2] A. N. Phillips, V. Cambiano, F. Nakagawa, A. E. Brown, F. Lampe, A. Rodger, A. Miners, J. Elford, G. Hart, A. M. Johnson *et al.*, "Increased hiv incidence in men who have sex with men despite high levels of art-induced viral suppression: analysis of an extensively documented epidemic," *PloS one*, vol. 8, no. 2, p. e55312, 2013.

[3] US Center for Disease Control, "HIV Testing in the United States." [Online]. Available: http://www.cdc.gov/nchhstp/newsroom/docs/factsheets/hiv-testing-us-508.pdf

[4] S. J. Little, S. L. K. Pond, C. M. Anderson, J. A. Young, J. O. Wertheim, S. R. Mehta, S. May, and D. M. Smith, "Using HIV Networks to Inform Real Time Prevention Interventions," *PloS one*, vol. 9, no. 6, p. e98443, 2014.

[5] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce *et al.*, "The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions," *International journal of communication*, vol. 5, p. 31, 2011.

[6] P. S. Earle, D. C. Bowden, and M. Guy, "Twitter earthquake detection: earthquake monitoring in a social world," *Annals of Geophysics*, vol. 54, no. 6, 2012.

[7] S. D. Young, C. Rivers, and B. Lewis, "Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes," *Prev Med*, vol. 63, pp. 112–115, 2014.

[8] S. D. Young, "A big data approach to hiv epidemiology and prevention," *Prev Med*, vol. 70, pp. 17–18, 2015.

[9] N. Thangarajan, N. Green, A. Gupta, S. Little, and N. Weibel, "Analyzing social media to characterize local hiv at-risk populations," in *Proc. Wireless Health 2015*, pp. 11–20.

[10] E. L. Murnane and S. Counts, "Unraveling Abstinence and Relapse: Smoking Cessation Reflected in Social Media," in *Proc. CHI 2014*. ACM, 2014, pp. 1345–1354.

[11] J. H. Fowler, N. A. Christakis *et al.*, "Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study," *Bmj*, vol. 337, p. a2338, 2008.

[12] M. A. Stoové and A. E. Pedrana, "Making the most of a brave new world: Opportunities and considerations for using twitter as a public health monitoring tool," *Prev Med*, vol. 63, pp. 109–111, 2014.

[13] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1, pp. 234–265.

[14] S. A. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in *Proc. SOCIALCOM '10*, Washington, DC, USA, 2010, pp. 88–95.

[15] C. Hutto, S. Yardi, and E. Gilbert, "A longitudinal study of follow predictors on twitter," in *Proc. CHI 2013*, pp. 821–830.

[16] J. L. Moreno, "Sociometry, experimental method and the science of society." 1951.

[17] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[18] R. Alun Jones, "The ethics of research in cyberspace," *Internet research*, vol. 4, no. 3, pp. 30–35, 1994.

[19] C. Kadushin, "Who benefits from network analysis: ethics of social network research," *Social Networks*, vol. 27, no. 2, pp. 139–153, 2005.

[20] B. Hoser and T. Nitschke, "Questions on ethics for research in the virtually connected world," *Social Networks*, vol. 32, no. 3, pp. 180–186, 2010.

[21] G. Eysenbach and J. E. Till, "Ethical issues in qualitative research on internet communities," *BMJ*, vol. 323, no. 7321, pp. 1103–1105, 2001.

[22] Z. Papacharissi, "The virtual sphere the internet as a public sphere," *New media & society*, vol. 4, no. 1, pp. 9–27, 2002.

[23] R. Priedhorsky, A. Culotta, and S. Y. Del Valle, "Inferring the Origin Locations of Tweets With Quantitative Confidence," in *Proc. CSCW 2014*. ACM, pp. 1523–1536.

[24] G. Seidman, "Self-presentation and Belonging on Facebook: How to Influence Social Media Use and Motivations," *Pers. Individ. Dif.*, vol. 54, no. 3, pp. 402–407, 2013.

[25] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You Are What You Tweet: Personality Expression and Eerception on Twitter," *J Res Pers*, vol. 46, no. 6, pp. 710–718, 2012.

[26] C. M. Rivers and B. L. Lewis, "Ethical research standards in a world of big data," *F1000Research*, vol. 3, 2014.

[27] J. Vitak, K. Shilton, and Z. Ashktorab, "Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community," in *Proc. CSCW 2016*.

[28] US Department of Health, and Human Services, "The Belmont Report - Ethical Principles and Guidelines for the protection of human subjects of research." [Online]. Available: http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report

[29] L. Hutton and T. Henderson, "An architecture for ethical and privacy-sensitive social network experiments," *ACM SIGMETRICS*, vol. 40, no. 4, pp. 90–95, 2013.