18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

A Network Science Approach to Open Source Data Fusion and Analytics for Disaster Response

Christian Anderson, Paul Breimyer, Stephanie Foster, Kelly Geyer, J. Daniel Griffith, Andrew Heier, Arjun Majumdar*, Danelle C. Shah, Olga Simek, Nicholas Stanisha, Frederick R. Waugh

MIT Lincoln Laboratory, Lexington, MA 02420
*Corresponding author: Arjun Majumdar (email: arjun.majumdar@ll.mit.edu)

***Abstract*—Network science is often used to understand underlying phenomena that are reflected through data. In real-world applications, this understanding supports decision makers attempting to solve complex problems. Practitioners designing such systems must overcome difficulties due to the practical limitations of the data and the fidelity of a network abstraction. This paper explores the design of a network science solution for the disaster relief domain with the goal of increasing the efficiency of disaster response efforts. Various real-world network science challenges are discussed relating to entity disambiguation and relationship estimation as well as general data science challenges such as limited access to representative data and learning inference models in this environment.**

A novel graph-based information management system was designed and prototyped to access and aggregate data from multiple sources. The system consists of five main parts: data ingestion, graph construction, inference, situational awareness, and evaluation. Data from open sources, such as social media, are ingested and fused to represent people, places, and social media users as a coherent social graph. This graph can be displayed to first responders to increase situational awareness or used as inputs to algorithms for graph analytics that support response efforts. Due to the lack of historical data from disaster events, an agent-based simulation was developed to create representative social graphs.

I. Introduction

No two disasters are the same, yet all response efforts face common challenges regarding the collection, processing, and dissemination of accurate and timely information. In particular, identifying and tracking the status and whereabouts of potential victims is critical for first responders, as well as for concerned families and friends who may be far from the disaster area. Efforts to manage and make sense of rapidly evolving incoming information are encumbered by the sheer volume of data originating from multiple distributed sources, often unstructured, and of varying quality and timeliness. A consequence of these challenges is that the time required to identify those affected by a disaster can extend to weeks or even months [1], [2]. This delay impairs relief efforts and places hardship on the family and friends of those affected.

This work is sponsored by the Department of the Air Force under Air Force Contract #FA-8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

The Semi-Automated Family Estimation (SAFE) system presented here offers technology solutions for improving the accuracy and timeliness of what is currently a very manual and time intensive process by using data from publicly-available sources. The primary goal of this system is to identify and determine the status of affected individuals. Challenges to accomplishing this goal include:

- ***Enormous data volumes***. For example, over 3.2 million tweets with hashtag #sandy were sent in the first 24 hours alone after Superstorm Sandy hit the United States in 2012.
- ***Heterogeneous, unstructured data sources.*** After the 2010 Haiti earthquake, relief workers identified the heterogeneity of available data sources (including text, geospatial information, photos, video, and social media updates) as a hindrance to their effective use [3].
- ***Unavailability of key data sources.*** A catastrophic event can make key data sources unavailable and can disconnect affected people from communications channels. For example, the Haiti and Japan earthquakes destroyed buildings containing police and school registries, important for determining who may be missing [3].
- ***Veracity of information and sources.*** Propagating incorrect information may be more damaging than propagating none. Misinformation about missing individuals intentionally propagated after the Japanese earthquake caused anguish to the families involved [4].
- ***Privacy.*** Aggregating information about people possibly affected by an event and making it available for analysis risks publicizing personally identifiable information.
- ***Analysis efficiency.*** Relief workers have many urgent tasks to accomplish in addition to identifying those affected. At the same time, the many volunteers who offer to help are often underutilized. [1], [3]
- ***Urgency.*** Speed in identifying those affected is critical not only to assisting them but also to mitigating the distress of family and friends.

As the amount of information available to relief workers has grown in recent years, technical solutions have been developed to address the challenges listed above. These solutions include the Ushahidi [5], Sahana [6], and QCRI [7] suites, university efforts such as Reunite [8] and ANPI-NLP [1], and

contributions from Google [9] and Microsoft [10]. Many of these – as well as other technologies not specifically tailored to humanitarian assistance – have been used effectively in response to a number of recent events. While these solutions provide a set of capabilities, none address all of the key challenges. The SAFE system aims to overcome many of these issues. This paper focuses on the ability to fuse a wide variety of publicly available data sources and perform inference on the data in support of disaster response resource allocation.

At the heart of the SAFE system is the concept that relationships among individuals are a powerful tool for inferring information about them. While other technologies focus on *lists* of individuals, SAFE represents the data using a *graph* that captures the many ways that people can be linked with each other. The nodes (or entities) in this multimodal graph are people, places, and social media users. The edges (or links) are relationships within modes such as "spouse of" and "friends with" or across modes such as "works at." The graph allows both visualization of the relationship information and provides a structure upon which analytics can be performed.

It should be noted that *privacy* is a very real concern for the disaster response community and for data scientists in general. In particular, as more of our lives are conducted or recorded online, there are ongoing debates within the government, industry, and academia concerning what the legal and technical requirements should be for accessing and sharing personal information [11], [12]. We have not explicitly tackled the issue of privacy here. All the data that serve as inputs to our system are publicly available and easily accessible via web browser and/or public APIs. However, it has been shown that the fusion of just a small amount of identifiable open data (e.g., tweets) can be fused with more sensitive anonymized data in order to perform strikingly accurate de-anonymization [13], [14], [15]. As this may be an unintended consequence of the fusion algorithms developed for our particular disaster response application, we have taken steps to mask the names and addresses of examples presented in this paper. Issues of privacy would need to be more thoroughly considered before a system such as the one presented here could be made operational.

The paper is organized as follows: Section II briefly describes the SAFE architecture. Section III discusses multiple data fusion algorithms designed for this system. A real dataset was collected to evaluate the performance of these algorithms individually. Section IV explores learning statistical models to support resource allocation in disaster management. To overcome the limited access to representative data an agent-based simulation was used to evaluate model performance. Finally, Section V provides a summary and presents ideas for future work.

II. SAFE System Architecture

The prototype system architecture for SAFE is shown in Fig. 1. It consists of five main parts: data ingestion, graph construction, inference, situational awareness, and evaluation.

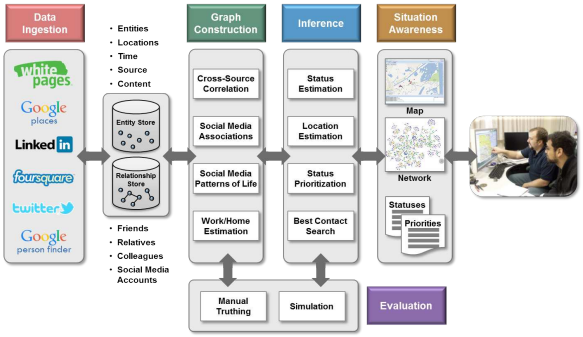

Fig. 1. Prototype SAFE system architecture

During data ingestion, information about residences and businesses in the affected area, as well as the people who live there, is collected from public records or ingested from online sources such as Google Places, Foursquare, and WhitePages. These data are collected along with activity from social media platforms like Twitter, Foursquare, and LinkedIn that provide additional information about where people work, where they visit, and who their relatives, friends, and colleagues are. On ingest disparate data is normalized into a common representation on a source-specific basis before being stored in a database and translated into a graphical representation.

The graph is constructed from both explicit links in the data and by estimating additional connections (e.g., a link to a person's work). During the course of the disaster response efforts various inference algorithms are applied to the graph data to estimate high-risk individuals and locations, prioritize search efforts, and identify paths of communication to potential victims. This information could be made available for access and modification by first responders or relayed to concerned friends and family.

III. Data Fusion

Fusion of data across multiple sources is the key enabling technology for constructing a single coherent graph, such as the network shown in Fig. 2. For example, unique locations ingested from Google Places must be fused with the corresponding locations ingested from LinkedIn (e.g., Soprano's Pizzeria). Additionally, while social media activity may be plentiful and rich, it is often necessary to associate a social media account user with a "real" person. Among the many data fusion sub-problems involved in constructing a multi-source graph, the following subsections highlight the location fusion and social media user to person association problems. Additionally, these algorithms are evaluated using a real dataset collected specifically for this effort.

A. Data Ingestion

Table I illustrates the common ontology used to represent data from a diverse set of providers. Open source data was collected from various social media and gazetteer portals accessible via the internet. Data was collected for a region

TABLE I
SAFE PROTOTYPE DATABASE

Database Entity	Description	Number of Records
People	Individual people. Fields include name, age, and status (e.g., BELIEVED_MISSING)	15,535
Locations	Geographic locations. Fields include address, geocoordinates, and location type (e.g., BUSINESS)	26,334
Social Media Users	Social media user accounts. Fields include username and platform (e.g., TWITTER)	74,324
Social Media Activities	Activities performed by a social media user (e.g., an individual tweet)	1,342,474
Graph Nodes	A node corresponds to an entry in the People, Locations, or Social Media Users tables	116,193
Graph Edges	A edge corresponds to a link between graph nodes. For example, a person and location may be linked by a LIVES_AT or WORKS_AT edge type	904,076
Graph Attributes	A key-value pair that provides a mechanism for defining generic attributes of graph nodes or edges.	1,084,680

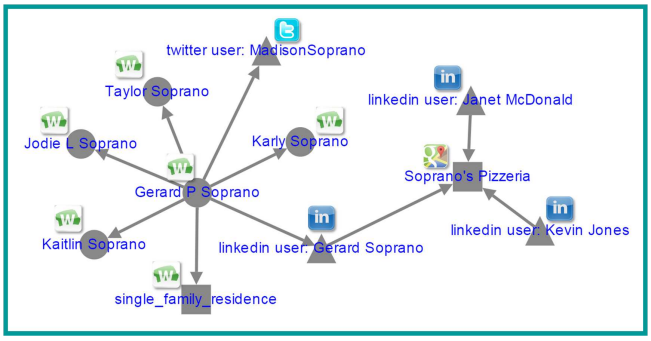

Fig. 2. Soprano family social network (with fictional names)

surrounding a mid-sized city and data providers were chosen to collect information about people, locations, and social media users. Table I illustrates the relatively large number of records that would be collected for each entity type and the resulting number of graph nodes, edges, and attributes. This dataset is used to provide qualitative analysis of the algorithms discussed in the rest of this section.

B. Location Fusion

Several publicly-available data sources include duplicate location information both within a single source and across multiple sources. The location names as well as any auxiliary information (e.g., location address) are often similar, but they are not always exactly replicated. Automated name matching algorithms have been developed and used for data disambiguation for several decades [16], [17], [18]. However, these methods are generally applied to person names (e.g., comparing "Steve Smith" to "Steven R. Smith") and do not account for auxiliary information. Accordingly, location fusion is framed as a supervised learning problem where given a pair of locations the goal is to classify if the locations represent the same physical location.

To train and test a classifier, 6,628 unique pairs of locations were manually classified into 4 categories corresponding to a human confidence that a pair represented the same physical location. The categories are as follows: 0) these places are not the same; 1) there is a nonzero probability that these places are the same; 2) these locations are probably the same; 3) these locations are definitely the same. An example of a category "2" might be two doctors that share the same building, but whose addresses ingested from Google Places list two different office numbers. An example of a category "1" might be two different stores in the same shopping mall.

Several features were selected to represent location information pairs. These included features to encode similarity between pairs of name strings and pairs of address strings by calculating the Jaro-Winkler distance [19], the Levenshtein distance [20], and the number of matching word tokens. Additional features included the distance in meters between locations and indicator features for matched location type, same data source, and less than 5 meter proximity. Using this feature representation, a random forest classifier [21] evaluated against the manually labeled dataset described above had an accuracy of 68% on a holdout test set.

C. Social Media to Person Association

Determining that Twitter handle @MadisonSoprano belongs to the unique person, Madison L. Soprano who lives at 287 Main Street (as a fictional example), can be extremely helpful for someone that is trying to determine her safety or whereabouts. Yet, associating physical people with their virtual aliases is nontrivial. To perform social media account association, a two-prong approach of name matching and location matching is used as described in the following subsections. This approach was taken because while a display name may be a strong feature for social media user to person correlation, it can often be the most noisy and difficult to disambiguate for common names. Similarly, home location is a strong indicator of account ownership, but the "home" of a social media user may be ambiguous or shared with other residents.

1) Name Similarity: For name matching, the social media account handle and/or display name is compared to the names of all "physical" people in the database (e.g., people ingested from Whitepages). Many traditional string matching techniques [22] were not sufficient for this problem, as media account names and Twitter handles in particular don't typically

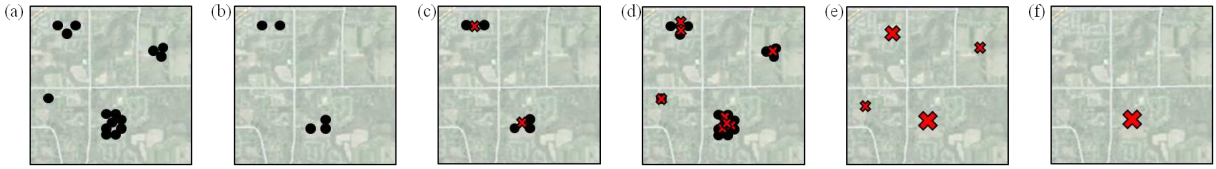

Fig. 3. Pictorial representation of home estimation algorithm for geotagged tweets: (a) activity filtered in time; (b) filtered activity quantized into temporal bins; (c) spatial clustering within bins; (d) clusters combined across bins; (e) spatial clustering of combined activity; (f) largest cluster selected.

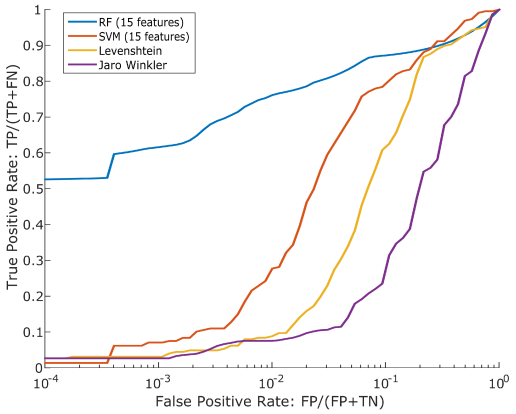

Fig. 4. ROC curve illustrating name similarity experiment results. The Random Forest (RF) classifier outperforms the Support Vector Machine (SVM) classifier using the described feature set. Additional, both methods outperform standard name matching techniques.

look like "FirstnameLastname." Rather, account handles often contain partial names, nicknames, initials, or no names at all.

A supervised machine learning approach was applied, similar to the approach described in Section III-B for location fusion. Fifteen features were calculated for every pair of person names and social media account names. These features included traditional string matching measures like Jaro-Winkler and Levenshtein distances, as well as features related to initials and gender. Gender can be a relevant feature if users use less formal but still common versions of their given name (e.g., Susan vs Suzie), if they are more likely to have a maiden name, and to help discriminate first name tokens that are not in fact names at all. First name tokens were compared to available US Census common baby names and ratios between male and female classes were calculated in instances where the name is common in both genders (e.g., Jamie). Additionally, last name tokens were compared to a list of over 151,000 commonly occurring surnames from the 2000 Census that accounts for 89.8 percent of the population [23]. The binary notation of real last name vs fake last name helps the model when words in social media names include objects or slang with a high string matching score to a person name, but are not a common surname. Over 90,000 pairs (corresponding to over 450 Twitter handles) were manually classified into four categories: 3) most likely the name person; 2) possibly the same person; 1) not likely the same person; 0) definitely not the same person, or no way to know. A random forest classifier [21] and a support vector machine (SVM) classifier were applied to learn a statistical model for associating social media accounts to people's full names. Fig. 4 shows performance curves on the manually labeled dataset described above for a derived binary classification problem constructed by grouping classes 0 and 1 and classes 2 and 3. The random forest classifier using the features developed for the SAFE system outperforms other algorithms; showing particularly good performance over the techniques designed for standard name matching.

2) Home Location Estimation: Name similarity is just one measure calculated for social media association. For accounts that also include geographic information, another similarity measure is calculated based on distances to estimated home and work locations. For sources such as WhitePages and LinkedIn, this information may be explicitly present in the data. For other sources, for example the small fraction (but still sizeable number) of Twitter users who geotag their tweets, the account's activity, clustered in both time and space, can be used to estimate likely home and work locations.

Note that most social media activity does not contain geocodes or other location identification (it's been estimated that between 80 to 98% of tweets contain no geographic metadata and only 1 to 2% include precise locations [24], [25]). There has been significant research in the area of estimating the home locations of non-geocoding Twitter users by leveraging their tweet content, tweet activity patterns, and social network [26], [27], [28], [29]. These methods have shown promising results, however the finest achievable location granularity is at the city scale, which is still too coarse for this application. For the work presented in this paper, home location estimation is only performed for users with location metadata.

Fig. 3 illustrates the home estimation algorithm pictorially: (a) first, all the media activity is filtered by time to consider the period of time when people are typically home (7 p.m. – 7 a.m. in the prototype system); (b) all filtered activity is divided into bins of length hour for the user's entire social media history – one hour bins balances general mobility patterns and non-random high social media activity periods; (c) for each one hour time bin, activity is clustered spatially using agglomerative clustering with a radius of 50m. This distance is fine enough to differentiate between houses and businesses fairly well, yet coarse enough to combine co-located activity subject to typical GPS errors; (d) cluster centroids for all one

hour time bins are combined; (e) centroids are clustered again spatially; (f) the largest cluster(s) are assigned as the social media account user's likely home location.

While a sufficiently large truth dataset has not been curated for this type of data, informal evaluations show the potential for the previously described technique. With a very small sample size of 58 users, this algorithm achieved 81% accuracy estimating users' home locations to within 10m and 90% accuracy within 50m. Finally, when the data is available available a comprehensive approach to social media users to person association could combine the name similarity and home location outputs simply as a weighted sum. A threshold could be applied to the combined output to limit the number of noisy edges added to the graph.

IV. Graph Analytics

Within the SAFE architecture, the graph analytics (or inference) component attempts to characterize how the impacts of a disaster are reflected in the social network. The goal is to learn a model that explains the observed effects while generalizing well enough to allow estimation on the unobserved population.

The primary use of the learned model is to prioritize disaster response efforts towards parts of the population that are most likely impacted by the disaster. A comprehensive solution to this problem would account for policy and implementation concerns that are outside the scope of this effort. Instead, this work will focus on a simplified abstraction of the resource allocation problem.

Data mining in this context is complicated by common social network analysis challenges: the data is large, noisy, and dynamic [30]. Additionally, each disaster is unique and its reflection in the social network will have its own distinct signature. Consequently, this work focuses on an online learning formulation of the problem where a model is trained and updated using data exclusively from the current disaster.

Using automated analytics in real-world scenarios requires the trust of the user community to facilitate technology adoption. A common approach in data science is to demonstrate performance on historical data. In scenarios such as disaster relief, where representative historical data is sparse and models must be generated on-the-fly, we propose that generating interpretable models is a reasonable alternative. Efforts will be made to explore learning algorithms that informally carry this property.

The final challenge in developing solutions in this space is the lack of labeled data to assess performance. Thus, we will first introduce an agent-based simulation approach that will be used to generate data to evaluate our learned models.

A. Agent-Based Simulation

One of the most challenging aspects of developing technical solutions for disaster response and humanitarian assistance is the unavailability of relevant datasets for evaluation. Typically, any data that is collected during a disaster event is permanently deleted within weeks of the event's conclusion. Moreover, the long-term storage of this data, even by groups hoping to conduct research, is either illegal or strongly discouraged as it becomes harder to take cognizance of the privacy of disaster victims as the data is distributed. Because of the lack of real data, a mixed-membership, agent based simulation model was developed to generate activity that was both statistically diverse and narratively sound. The simulation uses the same approach outlined in [31].

To create data representative of civilian day-to-day activity, four roles are defined: "home", "work", "public", and "available". Actions are defined as transitions from one location to another. The "work" and "home" roles have one possible action (reflecting the assumption that each agent has, at maximum, 1 work location and 1 home location), whereas the "public" role has many; one for each business location in the simulation. Agents in the "available" role have the full set of locations available to them. Simulations are run over one full day, split into 3 times corresponding to work hours (9 a.m. – 5 p.m.), evening hours (5 p.m. – 10 p.m.), and night hours (10 p.m. – 9 a.m.), with each agent drawing an event approximately every 60 minutes.

After a full list of roles and actions for every agent is generated, our approach differs from the approach explained in [31]. Specifically, the idea of a "social" action is introduced where two or more agents coordinate to meet at the same location at the same time. This is accomplished by first finding agents who are simultaneously in the "available" role, and choosing the agents with the most overlap to coordinate a social action. For an available agent, the number of additional agents that will participate in a coordinated social action is drawn from a Poisson distribution with mean 2. For the selected agents, the final meeting time and duration are computed by averaging the event time and duration of the overlapping available windows. The final coordinated action is drawn from a uniform distribution over the union of the sets of possible actions for each available agent. If no other agents are simultaneously available, the agent gets reassigned to a "public" action. The final result of this approach is that every "available" agent has been either reassigned to a "public" action or to the newly defined social action with one or more additional agents.

To increase the narrative consistency of the simulated data, a constraint is added stating that only agents in an "available" agent's one-hop network are eligible to participate in a coordinated social action with the "available" agent. To create the social network that informs the activity model described above, publicly available data is collected. Directly from these data sources, "lives with" and "lives at" edges are established between graph nodes representing people and locations. Then, using reasonable heuristics we create "spouse of", "works at", and "attends school at" edges. A generic "friends with" edge is randomly assigned to people who are connected through a location, because co-location is a predictive feature of social ties [32]; the probability of two agents being connected by this edge decreases exponentially with distance on the people-locations graph. Finally, a fraction of the people in the data are randomly sampled and assign them social media accounts. A

"social media friends with" edge is drawn between people who both own social media accounts and are otherwise connected on the graph. For this application, the ground truth of the social network is only important insofar as it affects the activity model. For that reason, the simplifying assumptions aim to generate consistent agent activity and are treated as truth in the simulation.

Once agent activities are generated that are consistent with the constructed social network, a disaster scenario is injected into the simulation. Disasters that occur rapidly (e.g., earthquakes and mudslides) are selected for this initial study. The effect of such a disaster is modeled by choosing an disaster event time before which agent activity is normal, and after which some area is affected.

B. Status Estimation

Following a disaster, information regarding the status of individuals within the potentially affected population becomes incrementally available from a variety of sources. For example, status information may be directly gathered using a tool such as Google Person Finder [9]. The current study formulates status estimation as a supervised learning problem where individuals can either be classified as affected or not-affected by the disaster. For a given point in time, a model is learned using training data from the subset of the population for which status is known. The model is then used to estimate the status of the remaining individuals and is updated as additional information becomes available.

With the problem expressed in a binary classification framework, classical supervised machine learning algorithms can be employed for model learning. Accordingly, a feature vector for each individual entity within the population is constructed to both capture entity attributes and properties of the individual's local neighborhood. A classifier is trained using samples contained within the observed subset of the population.

An alternative approach would be to adapt an algorithm from the set of collective classification algorithms developed for networked data [33]. The primary benefit of using a collective classification algorithm would be to directly model correlation with the unobserved neighbors of an entity of interest. This is accomplished by formulating a model that directly embeds a concept of label consistency. Applying these methods is nontrivial because learning algorithms in this online setting would need to address complex sources of bias such as high linkage and autocorrelation as described in [34]. Additionally, even simple collective classification algorithms (e.g., Iterative Classification [33]) add additional complexity, which makes the resulting model less interpretable. Thus, exploring the application of collective classification to this problem domain is left for future work.

1) Feature Engineering: At the onset of a disaster response effort, limited status information is available to train a model. This dictates the choice of both a simple model and a low-dimensional feature representation. Accordingly, a feature representation is designed to capture the local properties of the 1-hop network around a node or entity of interest. The resulting feature vector is composed of a set of four indicator features for each edge type listed in Table II. The indicator features represent if the node of interest is connected to one affected node, more than one affected, one unaffected node, and more than one unaffected node through the 1-hop network define by a given edge type. Using 14 edge types, this produces a 56-dimensional feature vector, which has been empirically observed to be sparse for real-world and simulated datasets. The sparsity is directly related to limited status observations and sparse network structure. In addition, for the simulated experiments described below a heuristic feature that created a 1-hop network from a combination of multiple edge types was manually defined to encode subject matter expertise (i.e. the knowledge of an experienced first responder). This feature was designed to represent human expertise that would be readily available in a real-world setting. While boosting performance, the heuristic did not dominate the learned model, demonstrating a non-trivial nature to the underlying phenomenology.

TABLE II
SAFE EDGE TABLE

Edge Index	Edge Name	Edge Index	Edge Name
1	attends school at	8	social media friends
2	classmates with	9	spouse of
3	friends with	10	teacher student
4	is near	11	uses account
5	lives at	12	visits
6	lives with	13	works at
7	related to	14	works with

2) Experiments: Experiments were run using the simulated datasets described in Section IV-A. For each dataset a disaster was simulated at one of three points in time (i.e. 12 a.m., 10:37 a.m. and 8 p.m.). Three binary classification algorithms were trained using the feature representation described above. These included the Bayesian Rule List (BRL) [35], Decision Tree (DT), and a linear Support Vector Machine (SVM). The algorithms were selected making considerations for model simplicity and interpretability. In particular, the BRL generates a decision list, which are a series of *if–then* statements that are highly interpretable by domain experts. On the other hand, the SVM has readily available, high performance implementations; making it an attractive choice when processing time is a concern. Empirical results demonstrate reasonable predictive performance indicating that the social graph indeed reflects patterns of the disaster that can be learned by a model. Additionally, results across the chosen set of classifiers are similar allowing a system level design decision between interpretability (i.e. BRL) and high performance (i.e. SVM).

C. Prioritization

While status estimation optimizes a model at a given instance in time, prioritization attempts to address the temporal nature of the disaster response resource allocation problem. The problem is cast in an active search framework where the

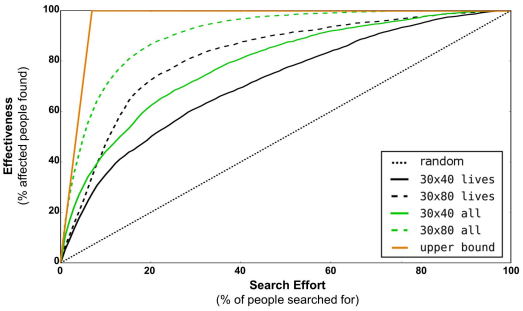

Fig. 5. Prioritization curves demonstrate the potential benefit of using social media data in disaster relief applications. The black curves indicate using only home location information and the green curves indicate using additional information from social media sources. The solid curves represent knowing 40% of the relationship information and the dotted curves represent knowing 80% of the relationship information. All curves begin with the status of 30% of population known.

goal is to maximize the number of affected individuals found [36]. Formally, given an initial state in which status information is observed for a portion of the graph, the performance of a search algorithm that incrementally selects a subset of the population to observe is measured. Information from these new observations can be used to adjust the search strategy as the search continues. The search concludes when the status of every individual in the population is known.

1) Experiments: To evaluate prioritization performance, the linear SVM described in Section IV-B was trained and iteratively updated as groups of individuals were observed. For each search iteration, the list of unknown individuals is ordered based on the confidence score output by the classifier and an observation group is selected from the top of the list. It was empirically observed that setting group size for model updates to approximately 3% of the population is an optimal operating point for our datasets. This behavior likely reflects an operating point balancing the predictive power of the model with exploring the population to support refining the model.

Results using the 10:37 a.m. dataset are shown in Fig. 5. In each experiment, 100 Monte Carlo trials were used to sample the initial state. The line color and style are used to denote the specific set of data sources and initial conditions used in each experiment.

a) Data Sources: The black curves represent a case in which only information about where people live is used. This information could be collected from "traditional" sources such as the WhitePages. The green curves represent a case in which additional relationship information is used such as peoples' work locations or social media connections. This information could be collected from social media sources such as LinkedIn or Twitter.

b) Initial Conditions: The solid curves represent initial conditions in which 30% of the node information is known and 40% of the edge information is known. For our simulated data this is the point at which a reasonable amount of information is available to learn a model. The dotted curves represent initial conditions in which 30% of the node information is known and 80% of the edge information is known. These initial conditions illustrate upper bound performance of using the data fusion techniques described in section III or simply collecting additional relationship information through other means.

As expected, the results in Fig. 5 demonstrates that performance improves with more relationship information either from additional data sources or through data fusion techniques. Specifically, after the first 20% of the search effort, the use of additional data sources provides over an 18% improvement in effectiveness. Similarly, overall performance, as measured by area under the curve (AUC), improves by over 15% when the amount of edge information is doubled from 40% to 80%. These results indicate that social media data has the potential to provide value in disaster relief applications.

V. Conclusions

In this paper, we presented a prototype system for ingesting and fusing multiple open sources of data to represent a population affected by a disaster as a social graph. Interacting with this graph can provide situation awareness to first responders, and help to identify how potential victims may be contacted via friends, family, and co-workers. Additionally, by understanding how a disaster is reflected in the social network, models can be learned to support prioritizing disaster response resources.

Leveraging open source data to support disaster response efforts presents many policy and technology challenges, several of which are addressed in this work. First, identifying and accessing appropriate data sources for any particular population poses a significant challenge, and ingestion is necessarily source-specific. For the system presented here, five sources (WhitePages, Twitter, LinkedIn, Google Places and Foursquare) were identified and data was ingested for areas surrounding a mid-sized city. In an operational setting additional data sources, including closed sources only available to government authorities, could easily be included in the framework presented.

The second significant challenge was fusing information across these heterogeneous sources into a coherent social graph. Specifically, two sub-problems were presented to illustrate the nuances of how problems in this space – in particular, name and location matching using social media data – diverge from existing literature. In both cases datasets were manually labeled to evaluate the performance of algorithms built on top of standard machine learning techniques. The promising results provide a basis for future work and motivate a need for standardized datasets to evaluate performance.

A third challenge addressed in this work is the need for significant amounts of "truth" data to evaluate solutions. With respect to disaster response, complete datasets generally do not exist. To address this concern an agent-based simulation was developed to generate representative social interactions at a large scale. Real data was used to seed the simulation and future work may explore extending this approach. Specifically,

using real datasets to validate the social interaction model and to design a network reaction model to apply at the onset of a disaster are two areas for future study.

Finally, we explored several inference methods for estimating status of unobserved persons in the social graph. This was framed as a supervised learning problem and classical machine learning techniques were applied and evaluated against simulated datasets. In particular, interpretable models were chosen in order to facilitate potential use in settings where non-expert users must be able to understand and trust the algorithm outputs. Performance on these data supports the hypothesis that relationships within a social network can provide useful information for estimating disaster effect phenomena. We expect that future work in this area may draw insights from the reinforcement learning community to marry the ideas of risk-aware algorithm search policies with public policy concerns.

REFERENCES

[1] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety information mining: What can NLP do in a disaster," in *Proc. 5th International Joint Conference on Natural Language Processing*, 2011.

[2] Oso mudslide. HeraldNet. [Online]. Available: http://www.heraldnet.com/section/osomudslide

[3] Harvard Humanitarian Initiative, "Disaster relief 2.0: The future of information sharing in humanitarian emergencies," UN Foundation & Vodafone Foundation Technology Partnership, 2011.

[4] "'Sick' messages falsely inform worried British families their relatives are dead after Japanese earthquake," Daily Mail, March 2011.

[5] Ushahidi Products. [Online]. Available: http://www.ushahidi.com/product/ushahidi/

[6] Sahana Free and Open Source Disaster Management System. [Online]. Available: http://sahanafoundation.org/products/

[7] Micromappers and Artificial Intelligence for Disaster Response. [Online]. Available: https://micromappers.wordpress.com/

[8] "Reunite case study: Harnessing computer intelligence to revolutionise disaster response," University of Manchester.

[9] Google Person Finder. [Online]. Available: https://google.org/personfinder/global/home.html

[10] Microsoft Solutions for Good. [Online]. Available: http://www.microsoft.com/about/corporatecitizenship/en-us/nonprofits/solutions-for-good/

[11] A. Stopczynski, R. Pietri, A. Pentland, D. Lazer, and S. Lehmann, "Privacy in sensor-driven human data collection: A guide for practitioners," *arXiv preprint arXiv:1403.5299*, 2014.

[12] C. Fuchs, K. Boersma, A. Albrechtslund, and M. Sandoval, *Internet and surveillance: The challenges of Web 2.0 and social media*. Routledge, 2013, vol. 16.

[13] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 2009, pp. 173–187.

[14] M. Korayem and D. J. Crandall, "De-anonymizing users across heterogeneous social computing platforms." in *ICWSM*, 2013.

[15] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the re-identifiability of credit card metadata," *High Impact Journal*, 2014.

[16] J. Liu, K. H. Lei, J. Y. Liu, C. Wang, and J. Han, "Ranking-based name matching for author disambiguation in bibliographic data," in *Proceedings of the 2013 KDD Cup 2013 Workshop*. ACM, 2013, p. 8.

[17] B. Lisbach and V. Meyer, "Name matching methods of the first generation," in *Linguistic Identity Matching*. Springer Fachmedien Wiesbaden, 2013, pp. 92–112.

[18] P. Christen, "A comparison of personal name matching: Techniques and practical issues," in *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*. IEEE, 2006, pp. 290–294.

[19] M. A. Jaro, "Probabilistic linkage of large public health data files," *Statistics in medicine*, vol. 14, no. 5-7, pp. 491–498, 1995.

[20] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] W. Cohen, P. Ravikumar, and S. Fienberg, "A comparison of string metrics for matching names and records," in *Kdd workshop on data cleaning and object consolidation*, vol. 3, 2003, pp. 73–78.

[23] D. L. Word, C. D. Coleman, R. Nunziata, and R. Kominski, "Demographic aspects of surnames from census 2000," *Unpublished manuscript, Retrieved from http://citeseerx. ist. psu. edu/viewdoc/download*, 2008.

[24] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook, "Mapping the global twitter heartbeat: The geography of twitter," *First Monday*, vol. 18, no. 5, 2013.

[25] C. Weidemann, "Social media location intelligence: The next privacy battle–an ArcGIS add-in and analysis of geospatial data collected from Twitter.com," *International Journal of Geoinformatics*, vol. 9, no. 2, 2013.

[26] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of Twitter users," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, 2014.

[27] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating Twitter users," in *Proc. 19th ACM international conference on Information and knowledge management*, 2010.

[28] H. Chang, D. Lee, M. Eltaher, and J. Lee, "@Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage," in *Proc. International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[29] S. Kinsella, V. Murdock, and N. O'Hare, ""i'm eating a sandwich in Glasgow": modeling locations with tweets," in *Proc. 3rd international workshop on Search and mining user-generated contents*, 2011.

[30] G. Barbier and H. Liu, "Data mining in social media," *Social network data analytics*, pp. 327–352, 2011.

[31] G. Bernstein and K. O'Brien, "Stochastic agent-based simulations of social networks," in *Proc. 46th Annual Simulation Symposium. Society for Computer Simulation International*, 2013.

[32] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proc. 12th ACM international conference on Ubiquitous computing*, 2010.

[33] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, p. 93, 2008.

[34] D. Jensen and J. Neville, "Linkage and autocorrelation cause feature selection bias in relational learning," in *ICML*, vol. 2. Citeseer, 2002, pp. 259–266.

[35] B. Ustun and C. Rudin, "Methods and models for interpretable linear classification," 2014.

[36] R. Garnett, Y. Krishnamurthy, X. Xiong, J. Schneider, and R. Mann, "Bayesian optimal active search and surveying," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012.