Towards Real-time Probabilistic Risk Assessment by Sensing Disruptive Events from Streamed News Feeds

Pete Burnap, Omer Rana
School of Computer Science and Informatics
Cardiff University, UK
(p.burnap),(o.f.rana)@cs.cardiff.ac.uk

Nargis Pauran, Phil Bowen
School of Engineering
Cardiff University,UK
(paurann),(bowenpj)@cardiff.ac.uk

***Abstract*—Risk management has become an important concern over recent years and understanding how risk models could be developed based on the availability of real time (streaming) data has become a challenge. As the volume and velocity of event data (from news media, for instance) continues to grow, we investigate how such data can be used to inform the development of *dynamic* risk models. A Bayesian Belief Network based approach is adopted in this work, which is able to make use of *priors* derived from a variety of different news sources (based on data available in RSS feeds).**

***Keywords*-risk modelling; streaming data and event analysis; dynamic data based modelling**

I. Introduction

All organisations, societies, infrastructures and ecosystems have at least one thing in common: their ability to operate, survive, or continue to function is dependent on a varying number of interacting and interdependent processes, actions and activities. This continuum has been defined as a ‘state of operation’ that exhibits different behaviours during normal operating conditions [1]. Risk to the ‘state of operation’ can be defined in two ways: things that could go wrong, a perception typically taken by risk analysts; and more recently the ISO 31000:2009 risk management standard redefined risk as ‘the effect of uncertainty of objectives’. This follows a more strategic and positivist viewpoint. In both cases risk can be considered as a casual factor influencing a ‘state of operation’ or more simply an *operational state*.

Risk is affected by dynamically occurring events. Today the risk to an organization’s operational state may be relatively low, but tomorrow an earthquake on the other side of the world, a freak storm, an economic downturn, an influenza pandemic, or simply a supplier going bankrupt, may lead to a significant disruption to the organization, causing its operational state to change. The chain of causality from a single event to a failed state of operation is often complex and the discipline of risk analysis aims to understand this complex set of interactions and dependencies, and manage them effectively. Disruption to one part of an operational system can have implications elsewhere and failures in one part of a ‘system’ can often “erode or overwhelm systems defenses elsewhere” through ‘risk pathways’ [2]. Take the example of the 2011 Tsunami off the coast of Japan where planning assumptions around the scale of a potential tsunami led to the construction of a sea wall of a given size. Once this line of defence was breached, the pumping sub-system has problems, which ultimately led to the release of radioactive gas. The uncertainty surrounding these events unfolded over time with media reports on the latest developments taking place in real-time [2].

It has been noted that risk to the operational state of early societies was local in its impact but that the modernization of society has increasingly created risk with global impact [3]. The globalization of business has led to a situation where, taking the example of the Tsunami in Japan, supply chains across the world were affected by the disruption to manufacturing in Japan. There are three key points to realise here: (i) a risk affecting one part of a system (e.g. a breach of the sea wall defense), has a causal impact on the likelihood of risks being realized elsewhere in the system (e.g. the ability to provide power to the pumping sub-system, and the ability to maintain a safe operational environment for the pumps to work), (ii) beyond the local system, the casual impact could have an influence on the operational capability of organisations on a global level due to the modernization of society and globalisation, and (iii) events unfold over time and as time elapses the likelihood of risks occurring changes, with information pertaining to these changes being reported with varying degrees of accuracy through various streams of information e.g. news stories, social media streams and from hardware sensors.

Given these three points, our research focus is therefore on building risk models that can support risk analysts in capturing the impact of single events and their causal effects on other parts of a system. Such risk models can be informed by using real time streaming data to update changes in the likelihood of risks occurring (based on reported news stories and social media posts, for instance). For example, if organization *x* based in Europe is critically dependent on organization *y* to supply components used in their manufacturing process – their ability to judge the importance of this relationship constitutes a risk model. Further, if organization

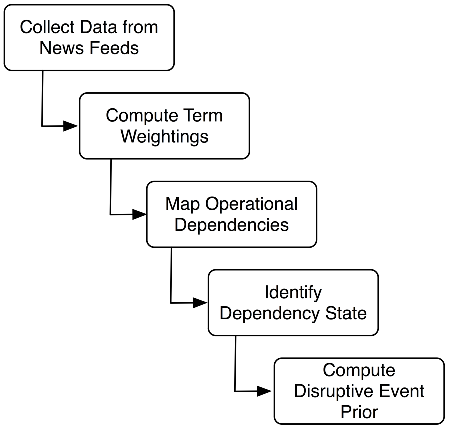
Collect Data from News Feeds
Compute Term Weightings
Map Operational Dependencies
Identify Dependency State
Compute Disruptive Event Prior

Figure 1. Event Analysis Process

y is geographically located in a high-risk earthquake zone such as in Japan or California, how can streamed media be used to inform *x*s risk model that a disruption to that dependency is imminent?

To achieve this we retrieve information over time from various streamed media sources that are published by news outlets of varying reputation and size. We analyze this information at specific points in time to create prior probabilities of event occurrence using a weighted theoretical framework based on confidence (in the source), freshness (recency of the information), and popularity (of the information across all information collected), which are then fed into a Bayesian risk model in the form of a Belief Net. The advantage of using Bayesian Belief Net is that they can be created very rapidly. Such models represent cause and effect through conditional probability, and are particularly suited to being updated with new information and therefore enable the expected outcome of the model to be generated given the new data. The latter point supports continual model monitoring and rapid reaction to new model forecasts as global events unfold.

II. METHOD

A. Sensing Events from Streamed News Media

Breaking news is reported in a number of *real-time* modes, of which some are programatically harvestable and lend themselves to the application of data mining techniques, such a content analysis and summarization. In this work we implement interfaces to a number of *Really Simple Syndication (RSS)* endpoints that provide a stream of text from local, national and international news broadcasters. Within the text are reports of events to which we apply a set of analytical techniques as summarised in the following steps:

1) Data Collection: For the purposes of experimentation we selected eight news sources of varying international size (e.g. BBC, CNN, Daily Mail, Wales Online etc). We developed a computational application to download the published RSS feeds of the sources every three hours. Each feed contained 10 news stories and, depending on the frequency of *newsworthy* events, there would be a number of new stories and a number already seen by our application in previous data collection runs. For each news story we save the title, description, Globally Unique IDentifier (GUID) and last publication date information into a database. We collected data in this way for seven consecutive days.

2) Term Weighting: Terms that appear in each news title are considered to represent the story. From the term set in each story description we first remove commonly occurring words (which reduce potential categorisation of the news story) by using a list of stop words. In order to determine the importance of each remaining term we calculate the Inverse Document Frequency (IDF) for the term. This is achieved by taking the total number of documents (news stories in the database), and dividing it by the number of documents in which the term appears – then taking the logarithm of this division. Having obtained the IDF, we calculate the Term Frequency (TF) by taking the number of times the term appears in the document in question, and dividing it by the total number of terms in the document. Using the IDF and TF scores, we can compute the TF.IDF weight of the term by dividing the TF by the IDF. This metric represents a measure of popularity of each term within a dataset. While we use a window of three hours, sliding time windows can be used to increase or decrease the size of the overall dataset (number of news stories analysed) when calculating TF.IDF and therefore distinguish between recent and older terms. Thus, a term with a high TF.IDF weight would occur frequently within a story, and in a high number of news stories, thereby highlighting the news as popular between news feeds and a story of possible interest during the previous time window. However, the relevance of the event is dependent on the effect it would have on an individual organisation, so we next need to determine the stories exhibiting terms of interest for previously identified risk factors.

3) Mapping Operational Dependencies: We use a Bayesian Belief Net (BBN) to represent a risk model (as described in Section B), where each node in the model includes a phrase that relates to an organisational dependency represented by that node. In this step we aim to identify news stories that are related to terms in these phrases. Suppose that the dependency is phrased as M4 Motorway open and Accessible or 'Clear airspace at Cardiff Airport'. To extract objects that could be affected by new events we can extract terms that are commonly named entities from these phrases (i.e. "M4 Motorway" and "Cardiff Airport") and configure our application to identify the presence of these terms in any of the news feeds in our database. Where the terms are present we proceed to determine the likelihood of a disruptive event that may affect this dependency by computing a probabilistic prior that can be fed into the Bayesian Belief Net (BBN).

4) Identify Dependency State: Disruptive events are by their nature counterproductive to the operational processes on which organisations are dependent. Thus, where we identify named entities within new stories, such as “M4 Motorway” or “Cardiff Airport”, we can further identify terms in the story that are synonyms or antonyms for the terms used to represent the dependency. For example, if a story reports an event involving “M4 Motorway”, then it stands that it will also include a term that represents its state (e.g. ‘open’, ‘closed’, ‘shut’, ‘blocked’, ‘locked’). If an antonym (opposite of ’open’) is identified we use the story as evidence of a likely disruptive event, and identifying synonyms indicates that the dependency will continue to remain in the required state. Data about the news story is then pushed into the next step, which is to compute a probabilistic prior.

5) Computing Disruptive Event Prior Likelihood: As we are deriving evidence-based event data and considering it as input for the likelihood of a positive or negative achievement of a dependent operation, we make use of Dempster-Shafer theory, which has been described as an appropriate method in assigning probabilities and dealing with uncertainty. Bloch discussed the key features of the theory in [4] and Beynon explained the advantages of the theory over those that have been proposed and used for decision modelling in [5]. In brief, the theory allows representing uncertainty in relation to an event by setting up an interval. The interval is based on two functions: Belief (B) and Plausibility (PL). The function B represents the confidence regarding the occurrence of an event, it is drawn from the sum of all the evidence that lead for such confidence:

$$B(E) = \sum_{A} M \tag{1}$$

Where E is the evidence supporting the event; A E and M is a mass probability which takes values in the range 0-1. During interpretation a higher value of translates as a high level of confidence towards the occurrence of the event. The function PL measures the extent to which we disbelieve the correctness of the event and is defined as:

$$PL(E) = 1 - B(E) = 1 - AM \tag{2}$$

where E is the evidence that contradicts the event. Thus, if the PL is 0.6 we state that the evidence that contradicts the event has a confidence of 0.4.

B. Modeling Dependencies and Calculating Impact of Events

To model risk we have used dependency modeling, which is a way of analyzing the risks to an enterprise [6]. The importance of the approach is to be explicit in how systems are configured such that a directed graph is developed containing a top-down model of dependencies. This approach is similar to a Fault-Tree or Failure Mode analysis – but rather than focus on what could fail, it addresses what is required to be operational. There is a difference here, and it is crucial. First, it considers a wide range of factors – not just technical. Second, by focussing on what could fail, it is possible to overlook potential failure modes that are not obvious. In cases where unforeseen failures occur (i.e. an unknown vulnerability is exploited), the failure mode perspective fails to assist in the analysis of how the system will react to this failure. Thinking from a goal-oriented perspective encompasses tacit knowledge of requirements for successful operation and, by underpinning the dependency model with computational conditional probability, quantifies the impact of the failure of one part of a system on all other parts in a stochastic model. The approach is based on the assumption that managers understand what they depend on every day to keep things running, but are not acutely aware of every possible mode of failure. It is based on the idea that risk is about goals and all risk springs from the fact that achieving our goals depends on many things, some of which we cannot control or predict or, in some cases, even understand. In this “dependability” context [6] defines risk as “the amount by which the probability of achieving our business goals is affected by things we cannot control, predict, or understand”, which aligns well with our proposed framework for sensing from streamed news feeds.

A dependency model is based on goals and objectives, and the prerequisites to satisfy these goals. In other words, it is a positivist, top-down approach working from goals to requirements. This is in strong contrast with other methodologies which focus on faults, disasters and failures (i.e., the “threat/vulnerability/impact” model). There are a number of advantages in the positivist approach, not least that it is easier and more intuitive to think of goals and requirements. Senior management people are more comfortable working with them than with disasters. It also allows “shared” goals between different departments in an organization’s dependency model, or even different organizations’ dependency models to be discovered and modeled as “shared risk”. As an example, suppose an enterprise has a daily requirement to transport supplies to and from suppliers and customers distributed across the country. Our goal is a successful delivery, and we are going to limit our analysis just to the transportation. The issues we are concerned with here are possible vehicular issues and possible travel problems. So, the success of our goal will depend on:

- A vehicle that works properly
- The availability of fuel
- The state of the traffic
- Possible road closures

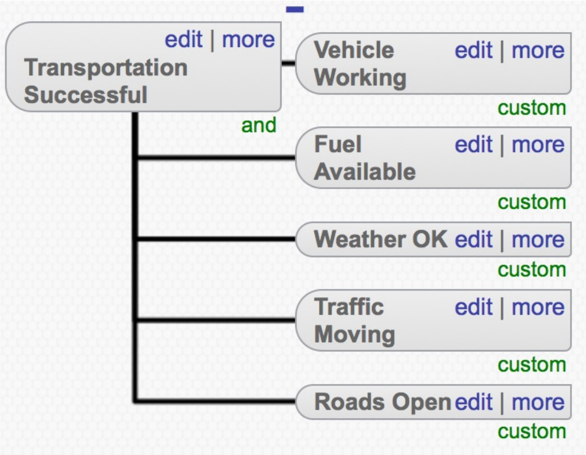

Figure 2. Basic Dependency Model

We will call these the dependencies of our goal. The list above is not supposed to be comprehensive, but illustrative of the dependency modeling approach. The success of our trip depends then on all these things, so we could draw a diagram as in Figure 2.

In a more realistic situation, the success of a goal would probably depend on many other items, but we use this simple scenario for the purposes of this discussion. We can view the entity ‘Transportation Successful’ as a goal with various degrees of achievability. It might, for instance, be fully achieved, partially achieved, slightly achieved, or totally unachieved, or any degree of achievement we care to come up with. For now, we only consider two extreme possibilities: failure and success. We can also view the dependencies – such as ‘Vehicle Working’, ‘Fuel Available’, etc as goals in their own right, similar to ‘Transportation Successful’. These too have various degrees of achievability and again for the moment we will limit the possibilities to just failure and success. It is important to note that these are not fundamental limitations, just convenient simplifications for the present discussion.

Elements on the extreme right of a dependency model do not have dependencies, or at least they are not shown in the model. These elements act as a sort of “given”. We can think of it as the point where risk and uncertainty enter the model. We can refer to such elements as ‘uncontrollables’ to emphasize that we can not do anything about changing their properties. However, they can represent a quantifiable metric that captures the probability of their success or failure. To support the computational calculation of probabilities in a dependency model we use a Conditional Probability Table (CPT). The probability of a dependant element being in its desired state (i.e. operational), given that another event has already occurred, is called a conditional probability and is the foundation of a Bayesian model. Therefore, the conditional probability that a dependant will be successful can be calculated from the given probability of success of its dependencies. A CPT specifies the probability of each state for a dependant for every combination of the states of all its dependencies. ’Uncontrollables’ have no dependencies within the model so we specify uncontrollables simply as a static probability. For example, we could state that for ‘Roads Open’, the probability that the desired state will occur (roads are open) is 90% or 0.9, while the probability of the undesired state is 0.1. This can be based on previous experience or existing data. With these given probabilities and the CPTs, it is possible to calculate the conditional probability of all the nodes in the model, right up to the ultimate goal of the model. In this paper we aim to derive the probabilities for ‘uncontrollables’ using a framework based on factors of the data obtained from streamed news feeds.

C. Towards a Framework for Calculating Input to Probabilistic Models

Events discovered from streamed media require some degree of ranking in order to determine the level of confidence that can be attributed to this kind of intellegence collected using automated methods. Sun [7], suggested a ranking scheme from events detected from data collected from the online social network Twitter, based on the following criteria:

- Confidence: A measurement of how much we trust the source (e.g. based on the reputation of a news source);
- Popularity: A measurement of how popular an event is in relation to all other detected events;
- Freshness: A measurement of how recent the event occured.

In this work, we have re-purposed this ranking scheme to enable us to derive probabilistic values for the ‘uncontrollables’ in a dependency model.

1) Confidence: The confidence of each of the news stories is represented using a *reputation* score between 0 to 1. A higher value indicates that the news story is collected from a highly trusted source, whereas a lower value indicates that the news story was published from a lesser trusted source, and may therefore contain inaccurate or untrustworthy information. The reputation scores are predefined in a manner such as displayed in Table I. The overall confidence of all news stories that are relevant to the success or failure of a dependant in our model is calculated as the mean confidence of all relevant stories, over all the stories collected via news feeds, and is is defined as:

$$CN_D = CN_{DS1} + CN_{DS2} + ... + CN_{DSN}/N \tag{3}$$

Where CND is the overall confidence for a collection of relevant news stories D and $CND1, ..., CNDsN$ are the confidence values for individual stories (s1,s2,....,sN) in collection D, and N is the number of stories in D.

Source	Confidence
BBC (bbc.co.uk)	1
Reuters (reuters.com)	1
Daily Mail (dailymail.co.uk)	0.8
Wales Online (walesonline.co.uk)	0.4

Table I: Confidence in News Sources

2) Freshness: The freshness of a news story represents a measure of recency and also takes a value in the range 0 to 1. This metric is derived from the time elapsed since the news story was published in the RSS feed, as shown in Table II.

Time (in hours)	Freshness
Below 2	1
2-4	0.8
5-7	0.6
8-10	0.4
Over 10	0.2

Table II: Freshness of News Sources

3) Popularity: The overall popularity of the event is simply defined as

$$Pop(E) = N1(D)/N2(D) \tag{4}$$

where Pop (E) is the popularity of the event in a collection of news stories D, N1 is the number of news stories containing both a dependency term (e.g. 'Motorway') and a term which relates to the success or failure of the dependency (e.g. 'Open' or 'Closed'); and N2 is the total number of stories analysed. The probability of the success or failure of an event is therefore formalised in a weighted equation as follows:

$$W1{*}confidence{+}W2{*}popularity{+}W3{*}freshness \tag{5}$$

Where W1+W2+W3=1 and W1, W2 and W3 are weights given to the confidence, popularity and freshness attributes, depending on the importance of each feature to the risk analyst. For example, one analyst may put more emphasis on confidence in a source, while another may place more importance on the recency of the story and therefore give more weight to the freshness value. Allocation of weights is therefore a subjective process.

III. Initial Results

For this example we focus on identifying disruptive events that may affect the dependencies in our model (see Fig. 2). We considered two dependency factors: 'Roads Open' and 'Weather OK'. Each dependency's state is binary, and hence, corresponds to two events. The prior probability of the belief (B) and plausibility (P) of these dependencies is formulated from news stories using the method discussed in the previous section.

a) For 'Roads Open' we use a term that relates to a specific motorway ('M4') and considers two outcomes (Open, Closed). b) 'Weather OK' represents the weather conditions on the transportation route at a particular time. The terms used are regional names along the route (i.e 'Cardiff', 'Reading', 'London') and weather terms ('rain', 'snow'). The outcomes for this factor are Good and Bad.

We assigned confidence values to eight online news sources, ranging from 0.9 to 0.5 and including national and local news providers. The values have been chosen to ensure that a story that arrives through a highly reputed source provides important evidence regarding any particular event. Further, we assigned the confidence, freshness and popularity weights with the values of conf=0.2, fresh=0.3,pop= 0.5, for formulating the priors of 'Roads Open' and 'Weather OK'

Table III presents the results of an analysis of news stories using the defined method on two seperate days. Based on the collection of the relating stories published on Day 1, the estimated plausibility (i.e. p(P) - likelihood of a disruptive event) is more indicative of an incident on the roads than Day 2. In the case of Weather, on both days there is no significant change in probabilities for the events corresponding. By treating the calculated plausibility p(P) and belief p(B) as priors, we can feed these metrics into our Bayesian Belief Net and recalculate the probability of success for each node.

	Roads Open				
	Conf	Fresh	Pop	p(P)	p(B)
Day 1	0.7	0.8	0.5	0.37	0.63
Day 2	0.75	0.6	0.67	0.34	0.67
	Weather OK				
Day 1	0.6	0.4	0.5	0.51	0.49
Day 2	0.6	0.4	0.5	0.51	0.49

Table III: Results

IV. Related Work

This work aligns closely with research that has been carried out for the forecasting of extreme events, such as weather events like hurricanes. Prediction of such events can be based on the development of complex physics-based models and computational simulations or, recently (and more relevant to this work) the development of network models that can be used to identify "community dynamics". A community is often identified as a common structure that can emerge in a variety of different types of systems – e.g. social networks, biological and weather/climate networks.

Stanhaeuser et al. [10] also identify these as data driven approaches for eliciting insights within complex networks and associated relationships. A key observation in such network-based models is that if a historic record of spatio-temporal events can be obtained, a supervised machine learning approach can be used to relate parameters that are used to characterise such events. Hence, occurrence of a set of events (observed with a certain degree of accuracy) could be used as a means to forecast the co-occurrence of another extreme event. Such a dynamic network-evolution model can be applied to a number of potential scenarios – and is not restricted to just forecasting of weather events. Often, as outlined in [8], the objective is not to predict an accurate numerical magnitude related to an extreme event, but to seek potential classification of an event into a number of pre-defined categories (i.e. provide enough information to facilitate a decision maker). Classifications such as "normal", "above normal" and "below normal" (for instance) can be used. The challenges identified by the authors also closely relate to this work, namely: (i) the multi-variate, spatio-temporal nature of the problem; (ii) the curse of dimensionality, i.e. the ability to deal with a very large number of features and identify those that are likely to impact the decision maker; (iii) inter-correlated and non-linear relationships, i.e. the occurrence of multiple operating "phases" of a system, with feedback loops existing between these phases. In [9], the authors have attempted to identified the occurrence of "anomalous" communities in such phased-based systems. Another perspective is taken by Rahwan et al. [11], where they consider how "social mobilisation" could be used to enable the detection of rare events (considered in the context of the DARPA "Network Challenge" [12] and the subsequent "Tag Challenge"). Both challenges aimed to leverage on social networks to locate 10 weather balloons tethered at random locations or required teams to locate and photograph 5 people across cities in two continents. The subsequent challenges of aggregating such information and ensuring that it is accurate through a verification process remain important challenges. Our approach is aligned with the methodology followed in this work, as we also attempt to find events that have a high information content and subsequently aggregate them using a Bayesian Belief Network. However, understanding how social data could be used to extend data from news sources (which already have undergone a verification process through an editorial team) would an useful addition to our work.

V. Conclusion

In this paper we have presented a framework for the detection of events from streamed news media, and the derivation of metrics that can be used with a probablistic risk model to translate the occurence of events into liklihoods of successful or failed operations within the systems of an enterprise.

Acknowledgment

This work was partly developed in the MSc thesis of Ms Thetida Zetta under the supervision of Dr. Pete Burnap.

References

[1] S Rinaldi, JP Peerenboom and TK Kelly, "Identifying, understanding, and analyzing critical infrastructure interdependencies", Control Systems, IEEE , vol.21, no.6, pp.11,25, Dec 2001

[2] Denis Fischbacher-Smith, "Destructive landscapes (Re)framing elements of risk?", Risk Management, vol. 13, pp. 115, 2011

[3] H Tsoukas, "Complex Knowledge. Studies in Organizational Epistemology", Oxford University Press, Oxford, 2005

[4] I Bloch, "Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account', Pattern Recognit. Lett., vol. 17, no. 8, pp. 905919, 1996.

[5] M Beynon, B Curry, and P Morgan, "The DempsterShafer theory of evidence: an alternative approach to multicriteria decision modelling, Omega, vol. 28, no. 1, pp. 3750, 2000

[6] The Open Group, "Dependency Modeling (O-DM) - Constructing a Data Model to Manage Risk and Build Trust between Inter-Dependent Enterprises". The Open Group, ISBN 1-937218-19-5, 2012.

[7] Y Sun, "Twitter Project" Online, Available at: https://wiki.engr.illinois.edu/display/dssi/2012+Twitter+Project.

[8] Zhengzhang Chen, Yusheng Xie, Yu Cheng, Kunpeng Zhang, Ankit Agrawal, Wei-keng Liao, Nagiza Samatova and Alok Choudhary, "Forecast Oriented Classification of Spatio-Temporal Extreme Events", Proc. of 23^{rd} Int. Joint Conf. on Artificial Intelligence, Beijing, China, August 3-9, 2013.

[9] Zhengzhang Chen, William Hendrix, Hang Guan, Issac K. Tetteh, Alok Choudhary, Fredrick Semazzi and Nagiza F. Samatova, "Discovery of extreme events-related communities in contrasting groups of physical system networks", Data Mining and Knowledge Discovery, September 2012.

[10] Karsten Steinhaeuser, Nitesh V. Chawla and Auroop R. Ganguly, "An exploration of climate data using complex networks". SensorKDD 2009 – Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data, pp 23-31. ACM Press.

[11] I. Rahwan, S. Dsouza, A. Rutherford, V. Naroditskiy, J. McInerney, M. Venanzi, N. R. Jennings and M. Cebrian, "Global Manhunt Pushes the Limits of Social Mobilization". IEEE Computer, vol. 46, no. 4, pp. 68-75, 2013. IEEE Computer Society Press.

[12] DARPA Network Challenge – details available at: http: //archive.darpa.mil/networkchallenge/. Last accessed: April 2014.