Virus Propagation On Social Networks Using Large-scale Biosensors

João Miguel Gonçalves de Sousa Andrade
Instituto Superior Técnico
joao.sousa.andrade@ist.utl.pt

Abstract—**The recent technological developments on mobile technologies allied with the growing computational capabilities of sensing enabled devices have given rise to mobile sensing systems that can target community level problems. These systems are capable of inferring intelligence from acquired raw sensed data, through the use of data analysis techniques. However, due to their recent advent, associated issues remain to be solved in a systematized way. Various areas can benefit from these initiatives, with public health systems having a major applicational gain. There has been interest in the use of social networks as a mean of epidemic prediction. Still, the integration between the mobile infrastructure and these initiatives, required to achieve epidemic prediction, is yet to be achieved. In this context, a system applied to epidemic prediction is proposed and evaluated.**

Keywords*—*Pervasive computing; epidemic prediction; large-scale sensing; social network analysis

I. Introduction

Distributed systems have been used as a platform to allow the interaction between groups of individuals and a set of devices. As technology advances in sensing, computation, storage and communications become widespread, ubiquitous sensing devices will become a part of global distributed sensing systems [1] [2].

Recently, the predominance of mobile phones equipped with sensors, the explosion in social networks and the deployment of sensor networks have created an enormous digital footprint that can be harnessed [3]. Furthermore, developments in sensor technology, communications and semantic processing, allow the coordination of a large network of devices and large dataset processing with intelligent data analysis [1].

The sensing of people constitutes a new application domain that broadens the traditional sensor network scope of environmental and infrastructure monitoring. People become the carriers of sensing devices and both producers and consumers of events [4]. As a consequence, the recent interest by the industry in open programming platforms and software distribution channels is accelerating the development of people-centric sensing applications and systems [4] [1].

People-centric sensing enables a different approach to sensing, learning, visualizing and data sharing, not only self-centred, but focused on the surrounding world [2]. These sensors can reach into regions, static sensors cannot, proving to be especially useful for applications that occasionally require sensing [5].

These systems constitute an opportunity for intelligent analysis systems, as relevant information can be obtained from large-scale sensory data and employed in statistical models [6] [1]. With these developments it is now possible to distribute and run experiments in a world-wide population rather than in a small laboratory controlled study [1].

Real-time user contributed data is invaluable to address community-level problems and provide an universal access to information, contributing to the emergence of innovative services [3] [2] [1]. For instance, the prediction and tracking of epidemic outbreaks across populations [3]. Thus, technological benefits are shifted from a restricted group of scientists to the whole society [2].

Healthcare is a possible application, where these systems can facilitate monitoring and sharing of automatically gathered health data [2]. Epidemics are a major public health concern and it has been shown impact can be reduced by early detection of the disease activity [3].

As most people will, in the near future, possess sensing-enabled phones, the main obstacle in this area is not the lack of an infrastructure. Rather, the technical barriers are related to performing privacy and resource respecting analysis, while supplying users and communities with useful feedback [1].

II. Related Work

A. Pervasive Computing

There is a tendency to augment devices with sensing, computing and communication functionalities, connecting them together to form a network, and make use of their collective capabilities [3].

Users become a key system component, enabling a variety of new application areas such as *personal*, *social* and *community* sensing. Each of these scenarios has its own challenges on how to understand, visualize and share data with others [2]. In community-level sensing, data is shared for the greater good of the community. In this area, social network analysis is a major source of information and relationships among groups of individuals. Applications only become useful once they have a large enough number of individuals participating. An infrastructure capable of integrating heterogeneous data sources is required, combining the resulting multimodal data and extracting behavioural patterns from it, through data analysis methods [3].

Part of this work took form under the Harvard-Portugal Collaborative Research Award HMSP- CT/SAU-ICT/0064/2009: Improving perinatal decision-making: development of complexity-based dynamical measures and novel acquisition systems.

An individual's role in these sensing systems may be *participatory* or *opportunistic* [3]. In *participatory* sensing, individuals decide which data to share, enjoying control over data privacy issues. In this approach, the target is restricted to a group of users willing to participate in the system [3]. In *opportunistic sensing*, sampling only occurs if requirements are met and it is fully automated, with individuals having no involvement in the data collection process [2].

The heterogeneity in data producers and information consumers leads to several challenges on data management [3]. The lack of correlation between data collected from distinct viewpoints and resolutions leads to an ineffective data merge and processing. To be able to integrate the system, data needs to be mapped to a shared vocabulary, respecting the same metrics [3].

These technologies are still in their beginning, leading to a lack of normalized architectures [1]. The placement of concerns on system components (e.g. remote servers, mobile sensing devices) has to be further researched [1].

Information needed by an application may only be available by integrating data from multiple sensing modalities [5]. Data analysis techniques require a systemic view, considering the sensing devices' resource constraints, communication costs to remote servers and the sampling rate required to detect and characterize interesting phenomena [2].

Some authors [2] [1] propose a three stage *Sense*, *Learn* and *Share* architecture.

In the *Sense* layer, sensing interaction-based mobility-enabled data is acquired from the heterogeneous sensors that are part of the system [2] [1]. Related applications may be present on the mobile sensing devices or remote server, communicating wirelessly [2].

In the *Learn* layer, information extracted from raw data is analysed using statistical measures, data mining or machine-learning techniques to infer higher-level meaning [2]. Data analysis techniques and features to analyse are chosen to best fit the availability and characteristics of the sensed data and the target application [2] [1].

In the *Share* layer, learned information is visualized and shared according to its application [2]. A personal application will inform its user and a community application will share aggregated information with its target group, while obfuscating their identity. Resulting information can also be used to persuade users to make positive behavioural changes [1].

B. Computational Epidemiology and Social Network Analysis

Computational epidemiology consists on the development and use of computer models to understand the diffusion of disease through populations with regard to space and time [7].

In order to accurately predict and understand the propagation of diseases, the data used in these models should be representative [8]. Nonetheless, decisions have to be made with limited information [9].

An epidemic model is a mathematical abstraction that describes the evolution of a transmittable disease in a population. Two of the most important notions in these models are those of the effective contact rate β, which stands for the rate of disease contraction, and the recovery rate δ, which is the rate of disease recovery.

It is relevant to distinguish between epidemics and outbreaks. An epidemic results from the spread of an infection from its initial set of cases to a community level, resulting in an incidence that has population-wide impact. An outbreak is associated with cases, whose transmissibility is inherently low. In this way, the infection dies out before reaching the general population [10].

The end of an epidemic is caused by the decline in the number of infected individuals rather than an absolute lack of susceptible subjects. Thus, at the end of an epidemic, not all individuals have recovered.

Social and biological systems can be described by complex networks whose nodes represent the individuals and its links the relationships between them [11]. Latest developments in epidemic spreading emphasize the importance of network topologies and social network analysis in epidemic modelling [11]. It involves the characterization of social networks to infer how network structures influence exposure risk.

Sampling matters in the creation of a credible mathematical basis for statistical inference on a social network. In sampling, the unit is the node, while the unit of analysis is most commonly the dyad or relation. The set of sampled nodes determines the set of sampled relations [12].

An important result in network models is the prediction of a non-zero *epidemic threshold* (λ_c). The higher a node's connectivity, the smaller the *epidemic threshold*, and consequently, the higher the probability of infection [11].

$$\frac{\beta_a}{\delta} \leq \frac{1}{\lambda_{1,A}} \tag{1}$$

Equation (1) represents the bounding of the *epidemic threshold* and defines a condition, that when not met implies the existence of an epidemic. $\lambda_{1,A}$ corresponds to the spectral radius, i.e. the maximum of the absolute value of the eigenvalues of the adjacency matrix associated with the contact network. β_a stands for the average rate of infection along a network edge and δ is the recovery rate of an infected node [13].

One way to accommodate the variety of contact patterns between individuals is by weighting the links of the contact networks. The weighted links distribute the contact rate parameter β_f over the network.

The weight value and its distribution can have a significant effect on the epidemic resistance of the topology, offering the possibility to alter a network without changing its topology. This introduction gives rise to a new form of clustering, i.e. weight clusters. Such clusters can boost infectious agent spread through the network [13].

$$\beta_a = \beta_f \overline{\omega} \tag{2}$$

Equation (2) introduces a weighting parameter $\overline{\omega}$ that accounts for the average contact rate placed on the network topology. β_f stands for the full contact measure of the effective contact rate β [13].

In these models, scale-free networks topologies are favoured. Their inherent large fluctuations between the number of connections in each node makes them appropriate to model real social networks [13] and computer virus epidemics [11]. In these networks, the final size and persistence time of a given epidemic are highly sensitive to the multi-scale hierarchical structure of the considered population [14]. For instance, nodes that are in contact with a large number of other nodes are easily infected and constitute a bridge for the spreading of infections [14] [15] [16].

C. Security

Respecting the privacy of its users is a relevant concern in a mobile sensing system [1] [3]. People are sensitive about how their data is captured and used. In the context of community sensing, there is the risk of leaking personal and community information. For instance, a connection between mobile sensors and observed parties may be implicit in their user's relationships [5].

Revealing too much context can potentially compromise anonymity and location privacy. Conversely, the inability to associate data with its source can lead to the loss of context, reducing the system's ability to generate useful information [5].

User control over data sharing allows users to define their participation in the system, empowering the decision making process [3]. One approach is keeping sensitive relations from being exposed, either by local filtering or by providing users with an interface to review data before it is released [5].

When mining social and community behaviours, anonymous data is needed [3]. Some approaches can help with these problems (e.g. cryptography, data and privacy-preserving statistics) [5] [1]. Nevertheless, they may be insufficient [1].

In opportunistic sensing schemes user trust may become a barrier to wide-scale adoption [2]. These issues may be addressed by providing sensing device users with a notion of anonymity.

III. SOLUTION

In an infectious disease it is necessary to detect, monitor and foresee the advent of an epidemic in a real-time environment. To operate in such a scenario the system should know who can get infected and which people have been in contact and where. Contact location, time and relationship with the subject are relevant metrics that affect the probability of disease propagation. Sensors and social networks analysis allow the integration of these concerns into personal devices, while developments in data analysis and modelling allow more accurate results regarding this data, potentially indicating community-level susceptibility to an epidemic.

This work comprises data gathering and management, intelligent analysis and privacy respecting sampling applied to epidemiological disease prediction in a population. A community is the target population of the analysis. It consists of the set of sensing devices belonging to people that are users plus their associated social contact network.

This solution considers performing robust data analysis in a dynamic environment and system scaling from a personal to a community-level, while providing useful feedback to its users.

The developed system is capable of exploiting large sets of multimodal data. Information extraction from raw data is done in a dynamic environment, by applying data analysis methods to the large-scale network data sourced from users and their sensed contacts. Intelligence is gathered in near real-time from a sensing network, in which only a population sample is considered.

These operations occur with some degree of distribution in the mobile sensing device and data processing backend. The criteria for this is based upon privacy, communication and resource management concerns.

Similarities between computer and biological infectious agents are exploited.

This work resulted in the papers: Social Web for Large-Scale Biosensors [17], Internet of Intelligent Things: Bringing Artificial Intelligence into Things and Communication Networks [18] and Epidemic Spreading Over Social Networks Using Large-scale Biosensors: A Survey [19].

A. Architecture

The proposed solution is composed by the architectural modules depicted in Figure 1.

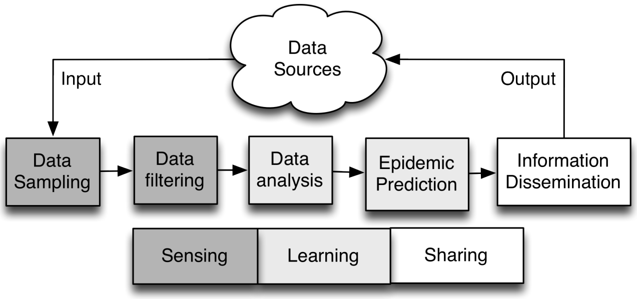

Figure 1. Solution Architecture

Both the *data sampling* and *data filtering* modules constitute the *Sensing* layer of the system. The aim of this layer is to obtain input. Sampling respects user privacy requirements, meaning that both a user's network and contact data are not disclosed to the system without explicit permission. Sensing, only occurs if requirements are met.

The *Learning* layer is constituted by the *data analysis* and *epidemic prediction* modules. In this layer obtained data is transformed and integrated into a model, contributing to the extraction of intelligence in the context of the applicational problem.

Information dissemination comprises the *Sharing* layer, where system output is returned to its users.

An analysis round constitutes a end-to-end system run, i.e. the flow of data from through these modules. From when it leaves the data source to the instant in which results are returned to the users. An actor is an entity that is part of a social contact network. All sourced actors have a corresponding system mapping, but not all of them carry sensing devices, i.e. they do not take part in the system directly. This means that not all actors are users.

An important distinction is the one between users and clients. Users are the individuals that participate in the system. Clients are the entities responsible for opportunistic data sampling.

B. Modules

1) Data Sampling: The purpose of this module is the acquisition of both network and contact data.

It is made possible through the use of mobile sensing device applications that are coupled with each user. Due to their limited coverage, the acquired data is inherently incomplete. There are different strategies that may be employed to tackle this. Network data sampling approaches, such as *ego-centric* sampling and the family of *link-tracing* methods can be attempted. *Link-tracing* methods can increase the network data covered by the system significantly. Alas, they come with a high computational cost [12]. Out of these methods, only *ego-centric* and *one-wave link tracing* remain reasonable in this aspect.

The information that is most relevant for the system and the one that can impact infectious disease spreading is user contact data. The acquisition of more network data nodes will not result in more data of this kind, as the system is already theoretically capable of obtaining the totality of the contact data existent in the social network platform.

In the light of these points, *ego-centric* sampling was considered to be the most adequate approach. While being subject to the same validation challenges as other methods, it yields objective friend connection results, it has a lower computational cost $O(n)$ per individual and all connections can be gathered from a single sampling request.

Finally, issues with missing data are mitigated by the connections observed in the acquired contact data. In this modality node relations for which there was no prior knowledge are sampled. These links also enable unobserved nodes to be detected. For instance, if a meeting involving a known node and a unknown node is detected, the system gains insight into missing node data.

Nonetheless, contact with strangers, i.e. nodes for which there is no connection in the data gathered for all modalities, is not accounted for. This constitutes a inherent limitation with the used sensing methods.

2) Data Filtering: This module has the goal of ensuring that unsuitable data does not get into the subsequent modules of the system. It functions as the safeguard before *data analysis*.

On the clients, it also functions as data anonymization module, ensuring that Facebook client identification does not reach the server. PBKDF2 is used to map this identification in a irreversible transform.

One drawback of this process is due to the fact that network and contact data, coming from different users, have to possess matching anonymized user identification. As such, it is necessary for different client applications to arrive at the same hash for data aggregation to be possible. This results in the need to have a common salt for all clients per analysis round. If an identification hash is obtained and this salt is discovered, an attacker will be able to compute the anonymized identification. Under these conditions, the present solution constitutes a compromise.

On the server, a part of this module's task is made possible through the use of programmatic checks. Due to the undirected nature of contact relations acquired for every user, sampling can result in duplicated contact data. The information to make this assessment is only available at this point in the server.

In the final filtering stage, each actor is represented by a network node that includes a unique identifier, reflecting user anonymity. The same is verified for each of its social connections. The data for perceived contacts is also included in the resulting data.

3) Data Analysis: The goal of this module is to join the data from different modalities (network and contact data) in a meaningful way and to output an epidemic model that can be used for the predictions of the next module in the chain.

A realistic network view, that can account for distinct forms of contact, can be provided by weighting network edges. Even if certain groups of individuals are topologically poorly connected, if the links between them have the highest weights, they become easily accessible by the infectious agent. Infectious agent access to pre-existing topological network clusters gets conditioned by edge weights, resulting in the creation of weight clusters [13].

This module takes advantage of these notions and encodes individual contacts in the weighting process applied to pre-existing network data. Thus, the sampled network data is extended with the notion of direct actor contact.

The operation process for this module is as follows. An actor name is the internal identifier used in data merging.

On the first stage the network and contact data are extracted for all actors that they reference, providing an universal mapping between an actor name and a matrix index i. This index refers to a given square matrix vector, where n is the number of the vectors and $i < n$.

On the next stage the network and contact data are transformed into adjacency matrices A and C. This is done using the mapping created earlier. As a result, both have the same dimension.

The next step introduces weights under the service of data merging. A_{ij} is a dichotomous variable from actor i to j, where 0 represents the absence of an edge and 1 its existence. As its upper bound is the number of contacts between i and j, C is restricted to $C_{ij} \in \mathbb{N}_0$. The rationale behind this is as follows. While contacts between two actors may increase over time, a social connection between them may either exist or not.

Taking matrix A, every non-zero entry is replaced by ω_{ij}, subjected to a weight bound $\beta_f \leq \omega_{ij} \leq 1$, creating the joint matrix W. W is the result of weighting matrix A with C.

$$\omega_{ij} = \beta_f^{\frac{C_n - C_{ij}}{C_n}} \tag{3}$$

The parametrisation for ω_{ij} is shown in Equation (3). β_a is the infectious agent effective contact rate. C_n is the total number of contacts for all nodes, while C_{ij} the number of contacts between nodes i and j and $C_{ij} \leq C_n$. Within the weight bound, edges that reflect at least a contact have their value scaled by a parameter that increases with the number of contacts that edge represents, forming a relative contact measure. This is based on the assumption that more contacts lead to a higher chance of contagion. Conversely, actors that have zero contacts between them are attributed β_a, as their relative importance in infection propagation is considered lower.

The final stage scales the joint matrix. As weighting directly decreases the spectral value and epidemic threshold of A, W has to be scaled. The scaling process consists in applying the average network weight to the matrix, substituting every W_{ij} with $\frac{W_{ij}}{\varpi}$, resulting in the matrix W'.

This aggregate dataset is a social contact network, which alongside with the infectious agent data, constitutes an epidemic model, where every sampled individual is explicitly represented. Its properties are described hereafter.

It is assumed that all individuals that are part of the social contact network are equally susceptible to infection. The spread of the disease is uniform in this aspect.

Model rates are constant over time. This means that in the execution of an analysis round they remain constant.

Vector and vertical transmission are not part of the model. Only direct contagion (horizontal transmission) between individuals is possible. Vector transmission is related to indirect contact, i.e. the transmission of Malaria with mosquito bites. Vertical transmission occurs between mothers and their unborn children.

The death and birth of individuals is not taken into account, resulting in no explicit demography. This assumption holds for fast diseases like influenza. For diseases that have a slower personal rate of evolution, akin to HIV, tuberculosis and hepatitis c, this is not valid, as during this length of time the population demography does not remain unaltered [20].

An infected individual is immediately able to transfer the infectious agent to others. Thus, there is no latent period of infection.

This epidemic model is passed to the next module.

4) Epidemic Prediction: This module's purpose is to apply the intelligent analysis algorithm to the merged data.

An analytic network topology approach was selected. The main idea behind the predictive power of this module is the notion that the topological properties of a social contact network can be used to assess epidemic persistence [13] and, thus, its advent. While this approach provides a limited context, it provides accurate results for the data family in analysis.

The prediction algorithm consists in identifying the spectral radius of the social contact network.

$$\lambda_{dom} \approx \lim_{n \to +\infty} (Tr(A^n))^{\frac{1}{n}} \tag{4}$$

Equation (4) exposes the formulation of the spectral radius identification method [21], where n stands for the iteration number and Tr is the trace of the matrix A.

Its computational cost of this method can be lessened by restricting the number of allowed iterations, provided that there is an approximation error that is suitable enough for the area of application [22].

The obtained spectral radius and the infectious agent parameters are inserted into equation (1), which relates the epidemic threshold of the graph associated to the network and the rate of infection transmission and recovery.

With this process it is verified if there is a subset of actors that possess enough connections, of a relatively high weight, to constitute a weight cluster, and thus form a network bridge. If this subset is representative enough, the existence of network weight bottlenecks will be high enough for the successful spread of disease and the creation of an epidemic.

The equations yields a boolean value, which can be translated into network epidemic vulnerability, if false, or the lack of risk if true.

5) Information Dissemination: Information dissemination comprises the communication of the prediction results, providing users with information that is both interpretable and useful.

Given that this work is targeted at epidemic prediction, the useful information at stake is the susceptibility of the community to a given epidemic, which is a binary decision, i.e. either a risk exists or it does not. By leveraging this metric, notifications that target the user community become viable, as the information to spread is depersonalized and is, thus, equally useful to all of the users.

C. Components

The system is composed by two distinct physical components: the mobile sensing devices (i.e. its clients) and the data processing backend (i.e. its server). The different logical modules or system concerns are distributed among them, according to Figure 2.

The communications inside the system are achieved with resort to Socket.IO [23].

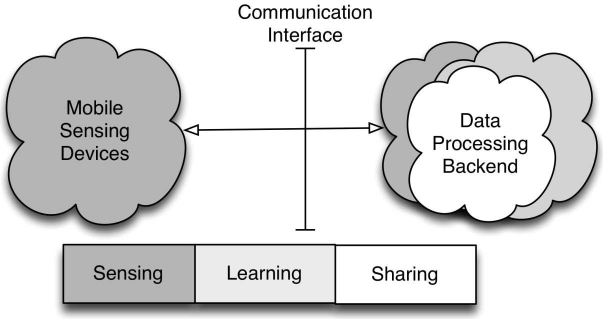

Figure 2. Module Distribution

The acquired raw data is originated in the *sensing* layer, located the mobile sensing devices, and is coupled to a client. Both network and contact data are sourced from the social network platform, i.e. Facebook. For privacy reasons, *data sampling* occurs entirely on the mobile sensing devices.

The data integrity component of the *data filtering* module is achieved on the backend, as only during the aggregation process it is possible to assess which data is likely to be erroneous and thus dispensable. The data anonymization part takes place on the mobile sensing devices, as they constitute a privileged position where is it possible to plug the leak of user identity.

Due to their complexity, the computation associated to the *Learning* layer takes place in a backend separated from the sensing devices.

The *Sharing* layer takes place in the backend, as by placing it there, relevant information can be easily acquired as output of the *Learning* layer. It can be transmitted to all the concerned individuals, as global knowledge required for client communication is already present in the system, while preserving privacy.

A criteria of ensuring there would be a minimal load on the mobile sensing devices was employed, by splitting the computational requirements and placing the demand on the backend.

A description of components ensues.

1) Mobile sensing devices: The mobile sensing devices are responsible for sampling data from the data source and for forwarding it to the data processing backend.

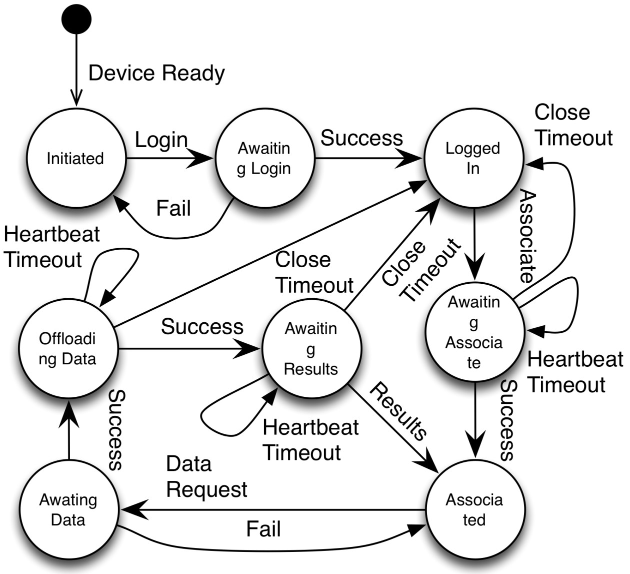

Figure 3. System client state machine

Figure 3 represents the state machine for the mobile client applications.

The client starts in the *initiated* state after the mobile device is ready.

A client may try to login to the social network platform and while it is in the *awaiting login* state, it may receive an access token, which represents a successful login. Any other reply or the lack of it, represents a failed login.

After the client is logged in, it has to associate itself with the data processing server. That is achieved with an associate message sent to the server.

While the client is in the *awaiting associate* state, it may succeed if it receives a message confirmation from the server or timeout. There are two kinds of timeouts. If a heartbeat timeout occurs, as the request is queued and the protocol will attempt to re-establish the connection to resend the message, the state is maintained. If a close timeout occurs, it implies that the connection is lost and that its socket was closed. As such, a new connection will have to be setup and a new message sent.

Upon a successful association process, the client will be *associated*. From here, it may receive a data request from the server, which will lead to sampling. A client may only hold data between server data requests. Upon this request the client attempts to re-sample its data from the social network platform and drops its previous state.

Now, the client will be *awaiting data*. The data may be either successfully loaded or the request may fail. Any response or absence of it that is not a message with the expected data format is considered a failure and the client will have to wait for the next data analysis round.

If the data is loaded with success the client will move to the *offloading data* state. Request success is assured upon server confirmation. Again both timeouts are possible. While a heartbeat timeout will lead to connection re-establishment retrials, a close timeout leads to the loss of the client connection socket on the server side and consequently to its disassociation. As no other server state is kept, the association process will have to start again.

The client will be *awaiting results*, until these are forwarded from the server. Heartbeat timeouts will lead to attempts to re-establish the connection. If that is not possible and a close timeout occurs, the association process will have to restart for the reasons referred before.

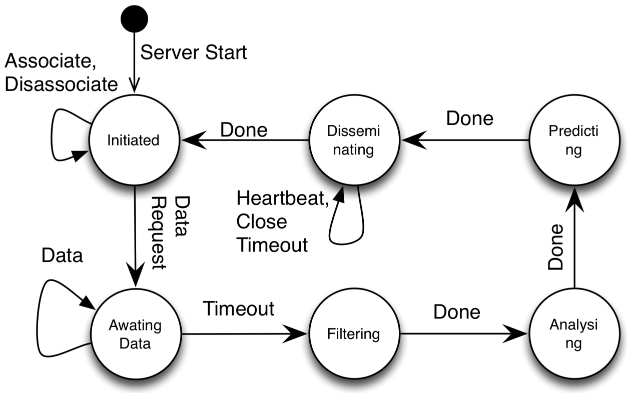

Figure 4. System server state machine

2) Data processing backend: The data processing backend is charged with filtering and analysing the data received from the mobile sensing devices. The results obtained from this process are sent back to the client devices. Its state machine is represented in 4.

After the server is started, it will arrive at the *initiated* state.

From here, it may receive associate requests from clients. In the same, clients may get disconnected leading to disassociate transitions.

Association requests are not verified by the server, implying that clients for data analysis are not authenticated. This leads to a trade-off between user anonymity and system security. While the system might be vulnerable to external entities that might offload fake data or attempt to overload the server, a given client cannot be verified without having access to its credentials. Furthermore, a user should not be asked to authenticate more than once. To solve this issue an authentication process between the clients and the server would have to be developed. Such approach would be based on the authentication of the mobile application, instead of the user itself. Finally, given the requirement of user anonymity, authenticating a given user is unfeasible, leaving room for abuse with non-authentic data. Nonetheless, since data is sourced from a third-party, any alike system is susceptible to such vulnerabilities.

After the analysis is triggered (either intentionally or within some pre-defined periodicity), the server will broadcast a data request message to all its associated clients. This set of clients constitute the partition whose data will be used.

Afterwards, the server will arrive at the *awaiting data* state. In this state client associations are no longer possible, as the server will receive data messages from the client partition established previously. If a given client does not offload its data within a predefined time, it is excluded from the partition.

When this timeout occurs, the server will start the *filtering* data state, resulting in the removal of erroneous values from user data.

The *analysing* data state ensues, originating a merging process of data with different modalities. Infectious agent data is integrated in this state. This results, in the combination of network and contact data with resort to graph weighting techniques and its parametrisation with infectious agent data.

The *predicting* state follows. The merged data is inputted to the epidemic prediction algorithm. After it is concluded, results are generated.

The server will start the *disseminating* state, where it will forward the prediction results to the client partition. Upon heartbeat timeout, the connection will be re-established and the results resent. On close timeout, since it results in the dismissal of the only link that exists between the server and its client, the results for the analysis round will not reach the client. After this state is left, all sourced data present on the server is dropped. Both these behaviours are a consequence of client privacy requirements.

IV. Result Assessment

Results are gathered through the execution of a set of experiments that were defined for each area of testing. Collected results are subjected to statistical analysis, by placing the gathered relevant summary statistics in a t-student distribution confidence interval. Finally, results are presented in a chart format.

The general assessment methodology focuses on testing, and if possible validating, the different parts of the system individually and then progressively integrating more complexity.

Figure 5 illustrates the general result assessment plan.

The learning layer and system-wide testing aim to validate the functionalities of the theoretical models.

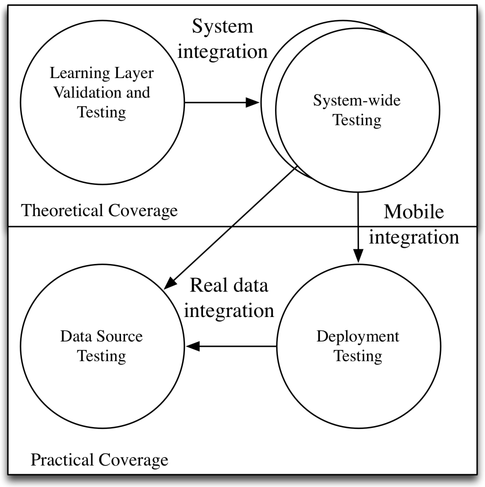

Figure 5. Result Assessment plan

Practical coverage modalities assess applicability of the system to a domain that is closer to the problematic. Deployment testing analyses the global architecture in a real mobile scenario with randomly generated data. Data source testing targets the system as a whole with real data for both local and mobile scenarios. Setting up the system to cover real data has the requirement of real users. Due to the sensitive nature of the system, the cost of deploying the system on a significant scale is high. Consequently, for validating the solution with a realistic data source context, this testing is performed with resort to Facebook's Test User API [24].

The objective of the following tests is the assessment of overall system end-to-end time. This metric accounts for both communication and processing time. The detailed experiment methodology is described under each test area subsection.

A. Deployment Testing

The network and contact data for this section were generated with resort to R, following respectively scale-free network generation methods and random sampling. These tests employ the use of real mobile sensing devices.

Under the same network, two Android smartphones were associated to the system. Their models are labelled in the charts, where appropriate.

The communication to the data processing backend goes through an access point.

A minimum of 50 timing samples were collected per combination of number of devices and network and contact data merges. The sample values were averaged and placed in a 95% confidence interval (CI). The sampling period is of $1500ms$.

A base network of 256 nodes was selected to be sampled by the users in the experiments of this section.

It was found that the differences in device capabilities affects the timing results. Moreover, device responsiveness was affected by the periodic activity of other running background applications.

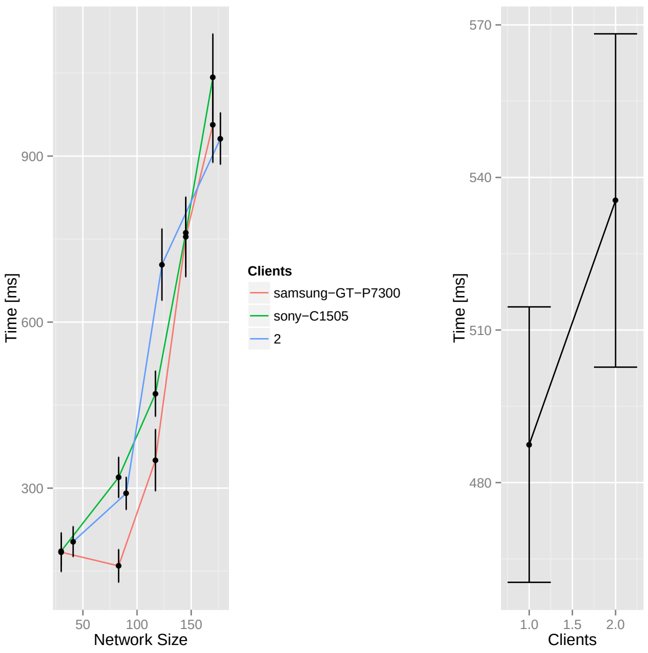

Figure 6. Real end-to-end time for network size and amount of users

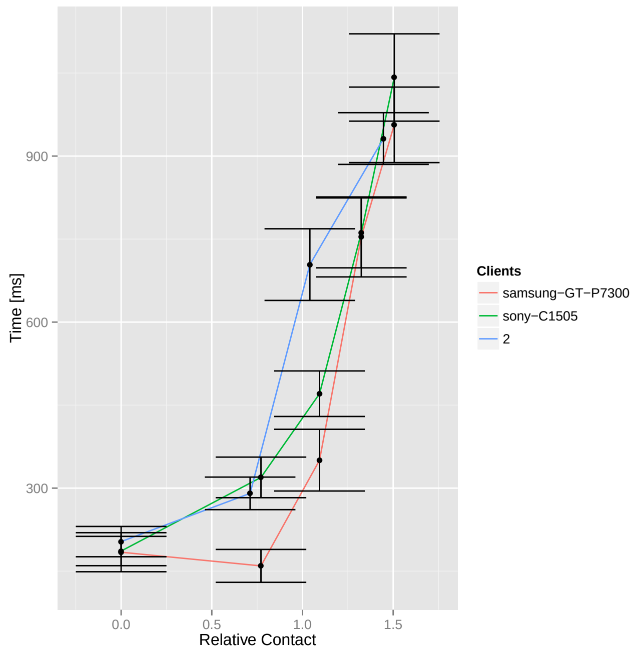

Figure 7. Real end-to-end time for contacts as a percentage of network size

1) Real end-to-end time for network and user base size: As shown in Figure 6, time evolves polynomially for the number of users in the system.

The major source of delay is the network size of the perceived network.

These delays also present a high variability. These results are according to what would be expected from a real communication environment, where packet loss and lack of connectivity are likely to occur. Also, differences in device performance affect the precision of the results.

2) Real end-to-end time and contacts: In this experiment randomly generated contact data is associated to the network data produced previously.

In Figure 7, it is visible that variable contact data impacts system response time. However, an increase in contacts is ultimately associated with an increase in network size.

The scalability of the system is constrained by the performance of its analysis algorithm. This algorithm is mostly constrained by network size, which is the factor that effectively influences end-to-end time.

The number of users also influences the delay in communications and it brings an additional cost in nodes sampled per user.

B. Data Source Testing

The aim of this section is to test the system with real data. To acquire data from a realistic data source, network and contact data were obtained from Facebook.

1) End-to-end time with a real data source: The purpose of this experiment is to assess the impact of adding a social network platform, as the system's data source, to the end-to-end time.

There are two scenarios to cover: local and mobile. For the local case, Facebook data accesses are achieved through a single machine, where the whole system runs. On the mobile scenario, they are done with resort to mobile sensing devices.

A base network of 15 nodes was chosen to be sampled in the experiment. 50 time samples were collected per each of the scenarios. These sample values were averaged and placed in a 95% CI. A single mobile device was used.

Figure 8 plots the end-to-end time for the two situations. It is noticeable that the overall end-to-end time is both higher and more prone to variation on mobile deployment. This is inherently related to the lower computational power of the mobile devices. Also the distributing of system clients over external devices has an additional level of communication and an associated delay.

The communication with the data source has a an associated cost in time, which is visible in comparison with Figure 6.

Either caching the data requested from it or receiving data updates in the form of events could contribute to the minimization of this factor's impact.

V. Discussion and Conclusions

This work presents an innovative system aimed at addressing the technical challenges of the multidisciplinary area that surrounds it.

The system considers the real-time processing of multimodal data, such as social network and location data from mobile devices. User data is anonymized and it is only shared after an initial consent, while subsequent sampling occurs

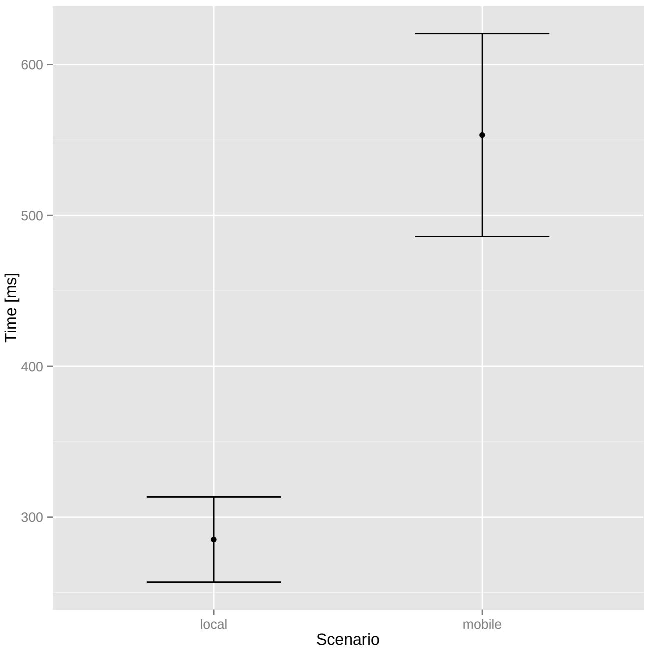

Figure 8. End-to-end time with a real data source

opportunistically. Such data is integrated with user social network data (originated from Facebook) and fed into a data merge and analysis algorithm for epidemic prediction.

The system is hence fully automated on an end-to-end perspective. Epidemic prediction is based on the analysis of the epidemic threshold of the sensed network and the infectious agent parameters, resulting in an aggregate metric for epidemic prediction. This analysis is centred on the sampled data and, as such, it is not generalisable to the whole population. Nonetheless, it sets the ground for algorithmic extensions for large-scale data merge and processing.

By delivering to users information, concerning outbreaks that might conduce to an epidemic, this system can contribute to the detection of communities that are vulnerable to a given infectious disease.

Finally, by weighting the concerns pertaining to the wide problematic spectrum of the area covered by this work, it was possible to arrive at a proof of concept solution that effectively implements these systems as a whole, targeting its problem domain with accuracy.

The solution deals with privacy concerns within some limitations. Nonetheless, complete data anonymization and analysis are two areas which pose conflicting approaches. As such, the techniques employed only go as far as feasible and should be regarded as a proof of concept in the system architecture.

More sensing modalities could be integrated into the sensing layer, providing a richer base of data to experiment with. For instance, the biosignal acquisition systems engineered during the Harvard-Portugal research could provide an extra dimension of information, possibly enabling epidemic models covering vertical disease transmission.

Contact data acquisition is focused on social network platforms. In line with the ground set for pervasive sensing, the perception of contacts through other means of lower location granularity and with a real-time data acquisition should be exploited. One example, would be the use of smartphone sensor access to sample GPS data.

The sampling process offered by the system can be improved in terms of the efficiency of communications. Client data can be sensed in real-time and cached until an analysis round is started. Presently, some social network platforms already provide support for event-based data acquisition in a publish-subscribe communication approach. If such a methodology would be employed, the redundancy of sampled data would be minimised.

The resulting system solves the issue of distributed resource management distribution by placing most of the concerns on a centralized entity, liberating the mobile sensing devices of unnecessary computation. Other approaches that compromise more may be superior, but their impact has to be properly assessed and studied.

The bottleneck to system inference is in the size of the sampled network. Through the path laid by this work, a more systematic analysis in the field of network data sampling should be raised, potentiating the impact of the currently achieved results.

Furthermore, significant performance improvements and system scalability can be unleashed by taping the potential of divide-and-conquer matrix analysis techniques and by extrapolating the validated analysis to the cloud.

Ultimately, the achieved architectural execution constitutes a stepping stone for other solutions, which may be aimed at a similar community problematic, but oriented towards distinct application areas. One example of such an area is the spread of computer viruses.

The principle underlying the executed systems approach is that the whole may be different from the sum of its parts. Consequently, it is legitimate that systems models outcomes may contradict the results that are originated from other reductionist approaches [25].

Predictive system results provides one level of inference, which does not conflict with other views provided by looking at the same data with different abstractions. Quoting Frederich von Hayek, the father of complexity theory, in his Nobel acceptance speech [26]:

"...as we penetrate from the realm in which relatively simple laws prevail [the physical sciences] into the range of phenomena where organized complexity rules...often all that we shall be able to predict will be some abstract characteristic of the pattern that will appear...yet...we will still achieve predictions which can be falsified and which therefore are of empirical significance".

REFERENCES

[1] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, A. T. Campbell, and D. College, "A Survey of Mobile Phone Sensing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 140–150, 2010.

[2] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, R. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G. Ahn, "The rise of people-centric sensing," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 12–21, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4557974

[3] D. Zhang, B. Guo, B. Li, and Z. Yu, "Extracting social and community intelligence from digital footprints: an emerging research area," in *Ubiquitous Intelligence and Computing*. Springer, 2010, pp. 4–18. [Online]. Available: http://www.springerlink.com/index/G85551H8L631837L.pdf

[4] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. Eisenman, X. Zheng, and A. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, 2008, pp. 337–350. [Online]. Available: http://dl.acm.org/citation.cfm?id=1460445

[5] T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich, "Mobiscopes for human spaces," *Pervasive Computing, IEEE*, vol. 6, no. 2, pp. 20–29, 2007. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4160602

[6] D. Peebles, H. Lu, N. Lane, T. Choudhury, and A. Campbell, "Community-guided learning: Exploiting mobile sensor users to model human behavior," in *Proc. of 24th AAAI Conference on Artificial Intelligence*, 2010. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/download/1933/2263

[7] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe, "EpiSimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks," in *International Conference for High Performance Computing, Networking, Storage and Analysis, 2008. SC 2008.*, 2008, pp. 1–12.

[8] L. Lopes, J. Zamite, B. Tavares, F. Couto, F. Silva, and M. Silva, "Automated social network epidemic data collector," in *INForum informatics symposium*, Lisboa, 2009, pp. 1–10. [Online]. Available: http://homepages.di.fc.ul.pt/~fjmc/files/workshoplopes-inforum2009.pdf

[9] P. F. Gorder, "Computational Epidemiology," *Computing in Science & Engineering*, vol. 12, no. 1, pp. 4–6, Jan. 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5372174http://darwin.bio.uci.edu/~kabbas/research/papers/bayes1h.pdf

[10] L. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, and R. C. Brunham, "Network theory and SARS: predicting outbreak diversity," *Journal of theoretical biology*, vol. 232, no. 1, p. 71.81, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519304003510

[11] R. Pastor-Satorras and A. Vespignani, "Epidemic Spreading in Scale-Free Networks," *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203, Apr. 2001. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.86.3200

[12] M. S. Handcock and K. J. Gile, "Modeling social networks from sampled data," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 5–25, Mar. 2010. [Online]. Available: http://projecteuclid.org/euclid.aoas/1273584445

[13] P. Schumm, C. Scoglio, D. Gruenbacher, and T. Easton, "Epidemic spreading on weighted contact networks," in *Bionetics 2007 2nd. Bio-Inspired Models of Network, Information and Computing Systems*, no. 1. IEEE, 2007, pp. 201–208. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4610111

[14] Z.-p. Li and G. Shao, "Halting Infectious Disease Spread in Social Network," in *IWCFTA'09 International Workshop on Chaos-Fractals Theories and Applications, 2009*, no. November 2002. IEEE, 2009, pp. 305–308. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5362017

[15] M. Kretzschmar, M. G. M. Gomes, R. a. Coutinho, and J. S. Koopman, "Unlocking pathogen genotyping information for public health by mathematical modeling." *Trends in microbiology*, vol. 18, no. 9, pp. 406–12, Sep. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20638846

[16] N. a. Christakis and J. H. Fowler, "Social Network Visualization in Epidemiology." *Norsk epidemiologi = Norwegian journal of epidemiology*, vol. 19, no. 1, pp. 5–16, Jan. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337680&tool=pmcentrez&rendertype=abstract

[17] J. a. Andrade, A. Duarte, and A. Arsénio, "Social Web for Large-Scale Biosensors," *International Journal of Web Portals*, vol. 4, no. 3, pp. 1–19, 2012. [Online]. Available: http://www.igi-global.com/article/social-web-large-scale-biosensors/75199

[18] A. Arsénio, H. Serra, R. Francisco, F. Nabais, J. a. Andrade, and E. Serrano, "Internet of Intelligent Things: Bringing Artificial Intelligence into Things and Communication Networks," in *Inter-cooperative Collective Intelligence: Techniques and Applications*. Berlin: Springer, 2014, pp. 1–37.

[19] J. a. Andrade and A. Arsénio, "Epidemic Spreading Over Social Networks Using Large-scale Biosensors: A Survey," in *Procedia Technology Volume 5*, 2012, pp. 922–931.

[20] M. Martcheva, "Introduction to Mathematical Epidemiology," 2001. [Online]. Available: http://www.math.ufl.edu/~maia/BIOMATHSEM/Lecture1.pdf

[21] E. Gonzalez, "Determination of the dominant eigenvalue using the trace method," *IEEE Multidisciplinary Engineering Education Magzine*, vol. 1, no. 1, pp. 1–2, 2006. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Determination+of+the+Dominant+Eigenvalue+Using+the+Trace+Method#0

[22] Z. Jelonek, "Solving polynomial equations," *Mathematica Aeterna*, vol. 2, no. 8, pp. 651–667, 2012. [Online]. Available: http://demmath.mini.pw.edu.pl/archive/dm45_4/4.pdf

[23] L. Labs, "Socket.IO: the cross-browser WebSocket for realtime apps," 2013. [Online]. Available: http://socket.io/

[24] Facebook, "Test User - Facebook Developers," 2013. [Online]. Available: https://developers.facebook.com/docs/test$\delimiter"026E30F$_users/

[25] A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea, "Social network analysis and agent-based modeling in social epidemiology." *Epidemiologic perspectives & innovations : EP+I*, vol. 9, no. 1, p. 1, Jan. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3395878&tool=pmcentrez&rendertype=abstract

[26] R. Rothenberg and E. Costenbader, "Empiricism and theorizing in epidemiology and social network analysis," *Interdisciplinary perspectives on infectious diseases*, vol. 2011, p. DOI:10.1155/2011/157194, Jan. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2992814&tool=pmcentrez&rendertype=abstract