Open Access

Open Journal of Internet of Things (OJIOT)
Volume 3, Issue 1, 2017

www.ronpub.com/ojiot
ISSN 2364-7108

Data Credence in IoT: Vision and Challenges

Vladimir I. Zadorozhny, Prashant Krishnamurthy,
Mai Abdelhakim, Konstantinos Pelechrinis, Jiawei Xu

School of Computing and Information, University of Pittsburgh, 135 N. Bellefield Avenue, Pittsburgh PA, USA
{viz, prashk, maia, kpele, jix20}@pitt.edu

ABSTRACT

As the Internet of Things permeates every aspect of human life, assessing the credence or integrity of the data generated by "things" becomes a central exercise for making decisions or in auditing events. In this paper, we present a vision of this exercise that includes the notion of data credence, assessing data credence in an efficient manner, and the use of technologies that are on the horizon for the very large scale Internet of Things.

TYPE OF PAPER AND KEYWORDS

Visionary paper: *Integrity, data credence, information assurance, data fusion, performance*

1 INTRODUCTION

In the next several years, we expect most environments to be "smart", in that there are data from sensors or actuators networked with wired and wireless technologies that are part of the Internet of Things – IoT. IoT has created a world in which tremendous amounts of data, pertinent to a specific application scenario, are being collected from disparate sources. The data generated by "things" will likely be augmented by crowdsourced data sources and context information that can be utilized for (a) rapid intelligent decisions on a variety of mundane and specialized problems, and/or (b) auditing and forensics to explain or understand a complex system that may have led to a spectacular event. An example of the former is the use of sensors by a bank for continually monitoring crop levels and soil moisture in land that belongs to a farmer who is using the land and crops as collateral for a loan [20]. An example of the latter is the unfortunate crash of a suburban train in New Jersey's Hoboken station [29].

In either case, we would expect the need to analyze data from a variety of sources -- data that may have intrinsic defects due to gaps in time and space (sensors located only in convenient areas or collecting data intermittently), fabricated or malicious data (due to adversaries), benign, yet erroneous data (measurement errors, cheaper sensors), data that have influence on the decision or forensics but may have been measured differently, with a coarser granularity (rainfall in the area around the farmer's land), or crowdsourced data (a tweet about the train's speed may provide an assessment of timing related to the accident). While the various data sources are related, their fidelity and reliability are highly varying. The *ground truth* of the data is not available, except that there is some level of trust in the devices that are gathering the data through the appropriate sensors or triggers.

This paper is accepted at the *International Workshop on Very Large Internet of Things (VLIoT 2017)* in conjunction with the VLDB 2017 Conference in Munich, Germany. The proceedings of VLIoT@VLDB 2017 are published in the Open Journal of Internet of Things (OJIOT) as special issue.

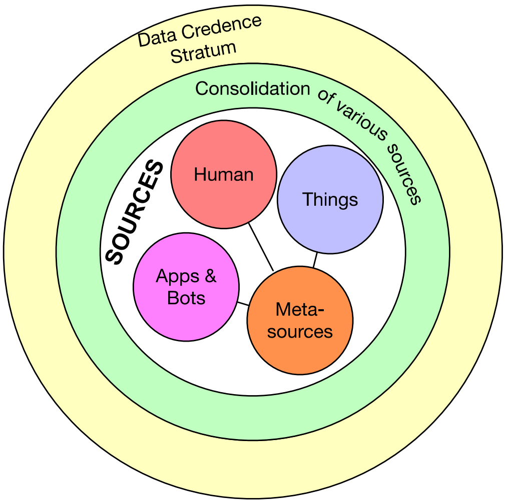


Figure 1: The data credence stratum

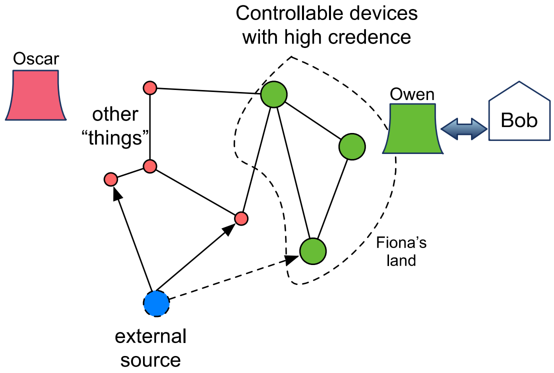


Figure 2: Example graph of types of and "links" between data sources

We expect that a central exercise in future environments will be assessing what we call *data credence*, which will provide a probabilistic range of confidence of the assessment of the data over time, allowing decisions to be dependable or audits to be trustworthy. Among the challenges that this assessment creates is the question of how we can assess and manage the credence in data in a general setting based on *a priori* and *evolving* confidence of data from the numerous smart "things" and the relationships between the reported quantities in different dimensions such as space, time, security, semantics, granularity, and context. Next, assuming this is possible, specific problems that then result in are how to perform such an assessment in an efficient way (in terms of data storage, latency, and energy) and how the data sources that can be controlled can be tuned to achieve the desired level of credence, if and where/when possible.

In this paper, we discuss this vision in general in Section 2. We consider a layered approach (see Figure 1), starting from what we call as the "credence stratum" in this section and show how this may work its way down to lower layers in later sections. In Section 3, we suggest the use of subjective logic and graph models as tools for assessing data credence. This section assumes honest data sources with benign or accidental errors. In Section 4 we discuss the challenges of fine tuning data credence that includes the use of cryptographic assurance of data. Section 5 looks at emerging technologies such as using multiple link layers and energy harvesting and how they may impact data credence. Section 6 concludes the paper.

2 DATA CREDENCE

The vision we present here is one where there is a determination of the credence of data in the emerging world of very large scale deployments of things that can monitor parameters and also actuate behaviors that influence the environment. We assume that data credence is a probabilistic metric (between 0 and 1) with a certain confidence. This maps neatly into the ideas of *subjective logic* as described in the next section. But first, we consider a specific (limited) example to explain the big picture.

2.1 Working Example of Bank and Farmer

Figure 2 shows a mix of three *types* of data sources: The size of the circle indicates an a priori credence (larger = better), the line stroke of the circle indicates the granularity (thicker = more granular, dashed = coarse), links indicate whether or not a source is "close" enough to another, perhaps along the dimension of space, and the color indicates the security level (green = authenticated, blue = external but credible, red = wild). Only for the purposes of illustration, let us assume that the data credence in this scenario is important to a bank "Bob" that is providing a loan to a farmer "Fiona" using her crops as collateral. Bob would like to verify whether Fiona is capable of repaying her loan and also to monitor the changes in this capability over time to make decisions regarding future loans or foreclosure.

Bob outsources this to a company "Owen" that deploys sensors in Fiona's land to monitor soil moisture, chemicals, and crop height. These would be the "green" sensors that have been authenticated by Owen. To save costs, Owen has deals with Oscar and Ogden who have deployed "red" sensors in neighboring land (small circles because their credence is lower). These deployments may gain or lose credence through verifiable reporting over time and what we call as meta-sources (see Figure 1), which may include communication patterns, locations, quality of devices, and so on. Further, the bank has external

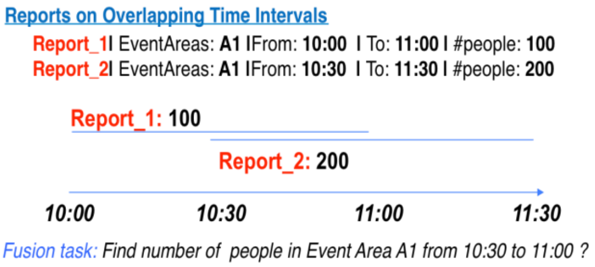


Figure 3: Example of data sources (event area reports) with overlap in time

knowledge from short and long-term weather reports and forecasts of crop viability from government reports (blue dashed circle - large because it is credible, dashed because it is coarse).

Data from these disparate sources have to be combined to provide a metric (perhaps a single value) that can inform Bob of Fiona's capability to repay loans (which is the decision to be made based on the credence of the data in question). This metric may be augmented over time with new data and by whether or not Fiona pays installments on the loan on time.

2.2 Sources of Credence and Disrepute

We can imagine that many sources add positively to the credence of data such as proximity and location, authentication, quality of measurements, previous reputation, and so on or on the contrary cause *discredit* to data. Consider an example of monitoring large outdoor events. This task requires large-scale information consolidation from heterogeneous data sources including infrastructure-based mobile systems, ad-hoc wireless networks and distributed Internet repositories.

Figure 3 shows an example of two population reports for event area *A1*. In this example, the data credence stratum obtains the reports from alternative information sources. *Report_1* is based on information estimated from surveillance cameras, while *Report_2* has been generated from a number of tweets posted by event participants and from communications within an ad hoc network of mobile devices of the event participants. Note that the reports estimate event population for overlapping time intervals with the total coverage from 10:00 to 11:30, and with overlap duration of 30mins. The task of the data credence stratum would be estimating the population dynamics within smaller time units (e.g., what was the most likely number of people in event area *A1* from 10:30 to 11:30?).

In some cases, analysis of relationships between overlapping data sources may reveal data inconsistency

Figure 4: Example of inconsistent data sources (event area reports)

that helps to assess the data credence. For example, several reports may reflect different numbers of people for the same location and time interval. Figure 4 illustrates a more complex case of inconsistency within four reports. The total number of R1 and R3 (550) should not be greater than the number reported in R2 (500). The number reported in R3 (250) should also be smaller than the number or R4 (200).

Consider a second example where a "thing" reports data to a sink about a phenomenon that it is monitoring. The data are sent at different times, some with cryptographic integrity checks and others without to save on computation and energy. The sink may attach more credence to data that has a verifiable integrity check and perhaps others that are close to it in time and content. In the latter case, the sink is looking at *consistency of data*, albeit in a different manner than that discussed in Figure 3. Credence here depends on verifiable integrity checks and proximity (defined with respect to the phenomenon) for those sources without integrity checks.

2.3 Data Credence vs Data Integrity

We argue that the notion of *data credence* is a superset of *data integrity* which is binary in nature. If there is (cryptographic) assurance that data came from the source from which they are supposed to have originated and that they have not been modified in any manner since then, we say that there is *integrity* of such data. If the cryptographic assurance fails, then the data cannot be trusted. However, in practice, it is virtually impossible (at this time) to ensure cryptographic integrity of all data, due to many factors. Such factors include technical ceilings on performance as well as issues such as cryptographic key management on the one side and policy, law, culture and human behavior on the other end.

Consider the issue of cryptographic key management. In a naïve setting, let us suppose that every message containing data from a "thing" has an integrity check using a (secret) key *k*. The challenges that arise are how we share the key and with whom. If the key is known to multiple entities, any one of them may modify the data without detection. From a network communications standpoint, the source "thing" and the destination (be it a gateway or the cloud) will perhaps share one or more keys. However, once the data reaches the cloud, the “ground truth” can no longer be verified. One may argue that digital signatures can be employed for this purpose, but this becomes computationally expensive if each message has to have a signature – both from the standpoint of signing and for verification of signatures for billions of messages.

In the meantime, it becomes necessary to use alternative techniques that evaluate the consistency of data as a means for credence as described next. To the best of our knowledge, using both data and cryptographic credence at the same time has not been previously investigated.

2.4 Related Work

The concept of data inconsistency and related concepts of data reliability have been explored mostly in the context of database management and data integration. The problems of data redundancy and inconsistency are of general applicability to large-scale Data Integration Systems. Data Integration Systems must address two major challenges: (1) *heterogeneous data* and (2) *conflicting data*. Resolving data heterogeneities has been the focus of active research and development for more than two decades [7] [19]. There are numerous tools for efficient mapping of data sources in a homogenous schema with proper data cleaning (eliminating typos, misspellings, and formatting errors), standardization of names, conversion of data types, duplicate elimination, etc.

The amount of research in the area of data conflict resolution and querying inconsistent data is also considerable. The work in [12] [5] [3] [4] provided a comprehensive review of the current state of the art. Early research on handling inconsistencies was mostly theoretical and did not relate this problem directly to the data reliability [23]. Data inconsistency as a key integrity constraint violation was considered in [1]. Consistent query answering that ignores inconsistent data, thereby violating integrity constraints, was introduced in [8]. This approach is related to more recent research on query transformation for consistent query answering [40]. An alternative approach is based on inconsistent database repair, producing a minimally different – yet consistent -- database that satisfies integrity constraints [38], [6]. Our work on information integration based on crowdsourcing and historical data fusion represent a new research direction in this area [28] [43] [45] [44].

Since the pioneering work of Grant [16] that first investigated the measurement of inconsistencies, in the past 20 years researchers have been trying to find the best way to measure inconsistencies. A good review of the research up to 2005 appears in [21]. Since then both additional inconsistency measures as well as properties that such an inconsistency measure should satisfy have been studied. The following are some of the important papers in this field: [22], [30], [31], [17], [18]. It turns out, as shown in [17] that the various proposed measurements are incompatible with one another, leading to the conclusion that the concept of inconsistency measure is too illusive to be captured by a single definition. So the best we can do is to find inconsistency measures that are the most appropriate in certain situations. Another issue here is that research on inconsistencies has been done primarily in an abstract setting using logic formulas. On the other hand, for the practical development of integrated systems researchers have used ad-hoc methods.

In case of multiple data sources a straightforward way to assess data reliability is to use a majority voting as a criteria for the most reliable data item. Meanwhile, reliability of data providers should also be taken into account, and some research has been conducted in this area. The first group of methods relies on probabilistic data accuracy assessment [11], [13], [27], [41]. Dong, et al. [11] proposed an *accuracy* technique, which calculates probability of each value being correct and averages the confidence of facets provided by the source estimating the provider trustworthiness. A more advanced *AccuracySimilarity* approach also considers the similarities of alternative values. Furthermore, [13] introduces a *POPAccuaracy* method assuming that false data value probability is uniformly distributed. A *TruthFinder* method, proposed by Yin, et al. in [41] differs from *Accuracy* by not normalizing the confidence score of each entity.

The second group of methods is based on web link analysis [32], [42], [14]. In [32], Pasternack, et al., proposed three techniques: (1) AverageLog is a transformation of Hub-Authority algorithm assessing source trustworthiness as an averaged confidence score of provided values multiplied by the log of provided value count; (2) Investment, where the confidence score of the value grows exponentially with the accumulated providers’ trustworthiness; (3) PooledInvestment differs from the i*nvestment* in that confidence score of data values grows linearly.

The work in [42] proposed a semi-supervised reliability assessment method called *SSTF*. This method assumes that there is a set of entities having true value affecting the result of the PageRank iteration procedure. The work in [14] proposed a *2-Estimates* transformation of Hub-Authority algorithm where provider trustworthiness is estimated as an average the vote count. They further proposed *3-Estimates*, which additionally considers the trustworthiness of data values. Other methods include IR-based techniques

[14], [34]. For example, Galland et al., [14] suggest maintaining a vector for a data value with each dimension corresponding to a provider. The reliability of the provider is assessed as a cosine similarity between provided values and selected reference values. In [34], Pochampally et al., proposed a method measuring the source precision and recall, and the correlation information between sources, based on which the confidence score of data value is computed.

None of the approaches discussed in this section take both inconsistency and assurance/integrity into account. In this paper we suggest an integrated data credence analysis exploring data redundancy and data inconsistencies so as to provide automatic data credence assessment that also takes into account issues of information assurance (through cryptographic methods). We consider an approach to discover data inconsistency through the analysis of relationships between data items and data sources with additional metrics that may include cryptographic assurance. Our work is the first attempt to utilize efficient inconsistency analysis and information assurance for implementing a scalable data credence stratum.

3 MODELING AND ASSESSING DATA CREDENCE

In this section, we start with the first of the many challenges that we discuss in subsequent sections – that of modelling and assessing data credence. *Subjective logic*, first introduced by Jøsang [24], combined with a *graph model* (as in Figure 2) that captures relations that may support or discredit the credence, appear to be best suited for this purpose. Recently, subjective logic has been applied for reliability assessment in both social and regular sensor networks [33] [36]. Combining subjective logic with flow-based reputation has also been explored in [37], which allows combining subjective logic with graph models. We suggest this as one of the approaches for modeling and assessing data credence, and illustrate this through an example below, although our examples use them separately.

Other approaches such as Bayesian inference, the Dempster-Shafer Theory of Evidence and Maximum Likelihood Estimation may be possible. The randomness in specific data sources may be characterized in some situations (e.g., location errors in GPS have known models) whereas it may be an assumption in other cases. Further, while ideally a joint probability distribution and the time variation in the case of stochastic processes would be the best for quantifying data credence (for example using confidence levels or outage probabilities – what is the probability that this data is correct within a specific range), most such analytical models are intractable unless sources are independent and processes are stationary.

Primary Challenge: How do we assess data credence?
We argue that traditional approaches for assessing data credence based on data consistency are insufficient in a world where we have disparate sources as described in Sections 1 and 2.1 with varying levels of a priori credence, much of which may be subjective. In this section, we assume that the sources are non-malicious and we relax this assumption later.

Assuming the data credence stratum continuously receives new data from multiple sources, it becomes necessary to determine credence values for (i) data items/reports and (ii) sources of these data, both of which evolve with the availability of new evidence. It becomes necessary to evaluate *internal credence* and *external credence* of data. It may be possible to use measures of "inconsistency" caused by a data source to assess its internal credence. While the assessment of internal credence can be a completely automated process based on objective metrics, it may be necessary in a human world to allow end-users to submit their subjective feedback on reliability of data and data providers to assess *external credence* (For example, how much trust would Owen put on Ogden Vs Oscar in Section 2.1?). With regards to the level of assessment, a *local* as well as a *global credence* may be necessary. The local credence value would be related to a single data item (e.g., report from a proximate thing), while the global one is related to a data provider/data source (Is the sensor from Ogden?). We explain these ideas next.

Internal credence: Handling internal credence requires solving the following two tasks: (1) finding efficient strategies to check for inconsistencies among data sources, and (2) finding the *least intrusive* inconsistency resolution strategy (this assumes sources are not malicious, but may be riddled with benign errors). For the first task, a considerable challenge is to optimize the inconsistency inference so it scales for large amounts of data. For the second task, we need to explore various minimal database reduction strategies to recover consistency. For example, in Figure 4 we can remove any of reports *R1*, *R2*, or *R4* and this will reduce the degree of inconsistency. Meanwhile, removing report *R3* eliminates inconsistency entirely, and thus represents the least intrusive inconsistency resolution. This also indicates a high probability for *R3* to be the least reliable report reducing its credence.

With respect to the *local (l)* and *global (g)* consistency, there exist interdependencies: It is likely that there will be a large number of the former that may be utilized to compute the latter. Local internal credence of a single data item/report may be assessed based on the inconsistencies with which it is associated, and then we may use a group of reports and their estimated credence to approximate the global internal credence of the data source. Each report may then be annotated with an internal credence tuple <*l*, *g*>, which includes the reliability for both the specific report and its source/provider. Maintaining the tuple may address the following problem: If we keep track of only the local credence, we may end up with a low credence value for a report that is actually accurate.

For instance, consider the two conflicting reports on number of people in an area A1 between 10:00 and 11:00. Report *R1* (from source *s1*) mentions 0 population, while report *R2* (from source *s2*) mentions 100 people. No other reports for this period are available. If we estimate local credence of the reports based on the inconsistencies caused by each of them, we obtain a local credence of 0.5 for both *R1* and *R2*. Meanwhile, the global credence of a source will reflect its accuracy and permit more adequate data assessment. Based on previous reports from *s1* and *s2*, we might reach global credence assessments of 0.3 and 0.8, respectively. Since their local credence is the same, we can use the global credence to increase or decrease our confidence in the data provided by each report.

External credence: We envision that users will be able to contribute their subjective data credence assessments on reports submitted either by themselves or by their peers. For instance, in the case where a contributor annotates her own reports with a reliability opinion, she might be aware of possible inaccuracies due to the method in which the data were obtained. Furthermore, she might be confident that the reliability of a conflicting report is low, due to her strong confidence in her own data. *There is a need to explore the challenges in external credence assessment and investigate a combined credence assessment.*

We next elaborate on the credence computation using Subjective Logic, as an example.

Subjective logic: Let *t*, *d* and *u* be non-negative values such that $t + d + u = 1$ and $\{t, d, u\} \in [0,1]$. Then, a triple $\omega = \{t, d, u\}$ is called an *opinion*, where components *t*, *d*, and *u* represent levels of trust, distrust and uncertainty. For example, high distrust with some uncertainty (10%) could be expressed as an opinion $\omega_1 = \{0.0, 0.9, 0.1\}$, while high trust with lesser uncertainty of (4%) could be expressed as opinion $\omega_2 = \{0.96, 0.0, 0.04\}$. By varying these parameters, we can express different levels of reliability in terms of which we can assess data credence. Subjective logic also provides a set of logical operators for combining opinions including conjunction, recommendation, and consensus. More details can be found in [24]. In our previous work [33], [36], we have successfully used subjective logic to express trust propagation in wireless sensor networks and in social networks. In this paper, we suggest its use towards scalable assessment of data credence.

Local internal credence: Every single source/report *r* can be deemed reliable or not with respect to its *degree* of inconsistency. It is possible to measure this degree through the percentage of *inconsistent conflicts* in which *r* is participating. For instance, let us assume that *r* reports on a time interval $[a, b]$ for data item *X*. Let us further assume that there are *k* reports that are related to data item *X* and their time interval partially or completely overlaps with $[a, b]$. Then if *r* is inconsistent with *m* of those reports, we can calculate its local credence (simple approach) as $LT(r) = 1 - \frac{m}{k}$. This calculation, however, provides a single point estimate for the local reliability, without considering the uncertainty of the assessment.

To assess the local credence of an opinion triplet, we can utilize the concept of *inconsistency level* (IL). We define an *inconsistency group* as a set of reports that have mutual inconsistencies. Then $IL(r, G) = 0$, if *r* is not a part of G; otherwise $IL(r, G) = \frac{1}{|G|}$. Hence, we can calculate *IL*(*r*, *G*), for each inconsistency group *G*. Next we can estimate a *mean inconsistency level* IL_{mean} of *r* and the corresponding standard deviation IL_{stdev}. We then define the local *distrust* on report *r* in the interval:

$$\left[\max\left\{0,\ IL_{mean} - \left(\frac{IL_{stdev}}{2}\right)\right\},\ \min\left\{0,\ IL_{mean} + \left(\frac{IL_{stdev}}{2}\right)\right\}\right].$$

Using a simple transformation, we can obtain the reliability of an opinion triplet. Assuming that the local distrust interval for *r* is [*y*, *z*], we have

$$w_1{}^n(r) = \{1 - \frac{y+z}{2} - \frac{z-y}{2}, \frac{y+z}{2}, \frac{z-y}{2}\}$$

as the local credence of the report *r* produced from source *n*.

To demonstrate assessing local credence of each report, consider the example in Figure 4. First, we need to find the inconsistency groups: G_1 = {R1, R2, R3} and G_2 = {R3, R4}. Then IL(R1, G_1) = 1/3, IL(R1, G_2) = 0, IL_{mean}(R1) = 0.167, and IL_{stdev}(R1) = 0.236. The distrust interval of R1 is [max{0, 0.049}, min{1, 0.386}]=[0.049, 0.386] and its local reliability opinion is ω_1(r1)={0.614,0.218,0.169}. Similarly, local reliability opinions for the rest of the reports are ω_1(r2) = {0.614,0.218,0.169}, ω_1(r3) = {0.465, 0.417, 0.118}, and ω_1(r4) = {0.573, 0.25, 0.177}.

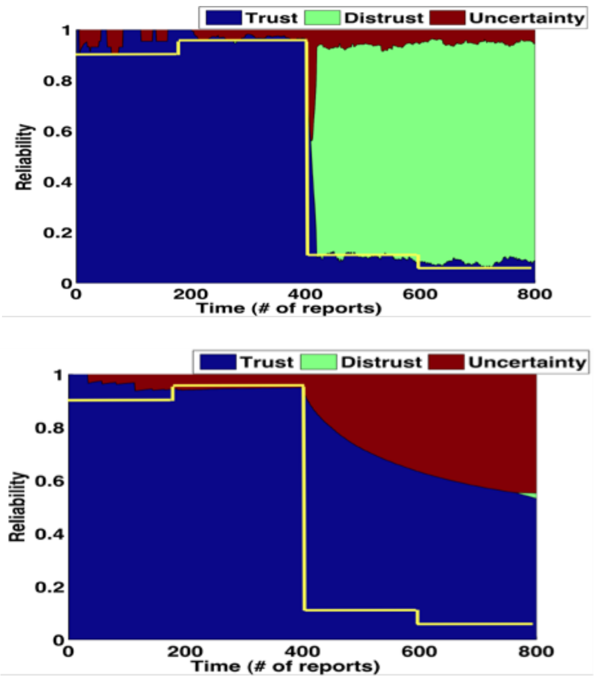


Figure 5. (top) Global credence in terms of reliability assessment: using the whole submission history; (bottom) Using the last 10 reports

***Global internal credence*:** After obtaining local reliabilities of each report, we assess the global credence of their provider *P*. Each report *i* submitted by *P* is accompanied by a local reliability opinion ω_i^n obtained through the process described above. To assess the global credence, we propose to aggregate them using the consensus operator of subjective logic [24]. We assume that every report is an agent, reporting on the reliability of its provider through the local reliability opinions. In brief, if $\omega_1^n(i) = \{t_i, d_i, u_i\}$ and $\omega_1^n(j) = \{t_j, d_j, u_j\}$, then their consensus is the opinion:

$$w_{i,j}{}^n = \{\frac{t_i u_j + t_j u_i}{k}, \quad \frac{d_i u_j + d_j u_i}{k}, \quad \frac{u_i u_j}{k}\}$$

where $k=u_i+u_j-u_iu_j$. Let us assume that provider Jack submitted all of the reports considered in the previous paragraph. Above, we calculated the local reliability opinions for all of the reports. The consensus of the opinions gives as the global credence opinion for Jack, which in our case is: $\omega_{r1,r2,r3,r4}{}^{Jack}$ ={0.629, 0.328, 0.044}. An interesting observation is that the consensus operator considerably reduces uncertainty: The more opinions we combine, the more *certain* the global opinion. Notice that the consensus of only the first two reports results in uncertainty of 0.092. Adding the third report reduces the uncertainty to 0.055, while the consensus of all 4 opinions results in the lowest uncertainty of 0.044.

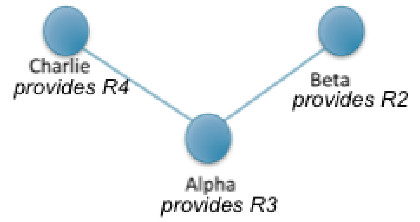


Figure 6. Example of an Inconsistency Graph

Figure 5 shows a *simulated* internal global credence assessment for a data source which provided 800 reports. The yellow line reflects the real user's reliability over time used in the simulations. In the left figure, we use the complete set of reports from the provider to assess her/his reliability, while on the right one we use only the 10 most recent reports. As we can see in the latter case, we are able to react to the reliability dynamic much faster.

***External credence*:** To handle external credence assessment, the rating process enables users to provide an external reliability opinion on data. Here, they can also provide external opinions on the global reliability of a provider (e.g., if they are aware of faulty data gathering etc.). These external opinions can be fused using the recommendation operator of subjective logic. When a user (say Jack) provides an external reliability opinion (local or global) ω_{ext}, the system can use its own opinion on Jack's external reports in conjunction with ω_{ext} to obtain a final external (local or global) reliability recommendation.

Another approach that we propose towards assessing data credence is to represent conflicting data in the form of an Inconsistency Graph and to use efficient graph analysis techniques (e.g., based on a modification of the well-known page-rank algorithm). We can generate the Inconsistency Graph (IG) with nodes corresponding to different data sources or data items and edges reflecting inconsistencies between the data source/items. Graphs with higher connectivity correspond to data with lower credence. For each node, higher connectivity means lower credence/reliability. Inconsistency with less reliable nodes is less severe than inconsistency with more credible nodes. Disconnected nodes correspond to data sources/items with the highest credence.

Figure 6 shows an example of an IG reflecting conflicts among three data sources providing event area reports from Figure 4. Here we assume that data source *Charlie* provides the report *R4*, while *Beta* provides *R2*. Both *Charlie* and *Beta* conflict with the data source *Alpha* providing the report *R3*. Note, that the conflict between *Alpha* and *Charlie* is more severe than the conflict between *Alpha* and *Beta*, since *R3* contradicts *R2* in combination with *R1*. Both number of conflicts

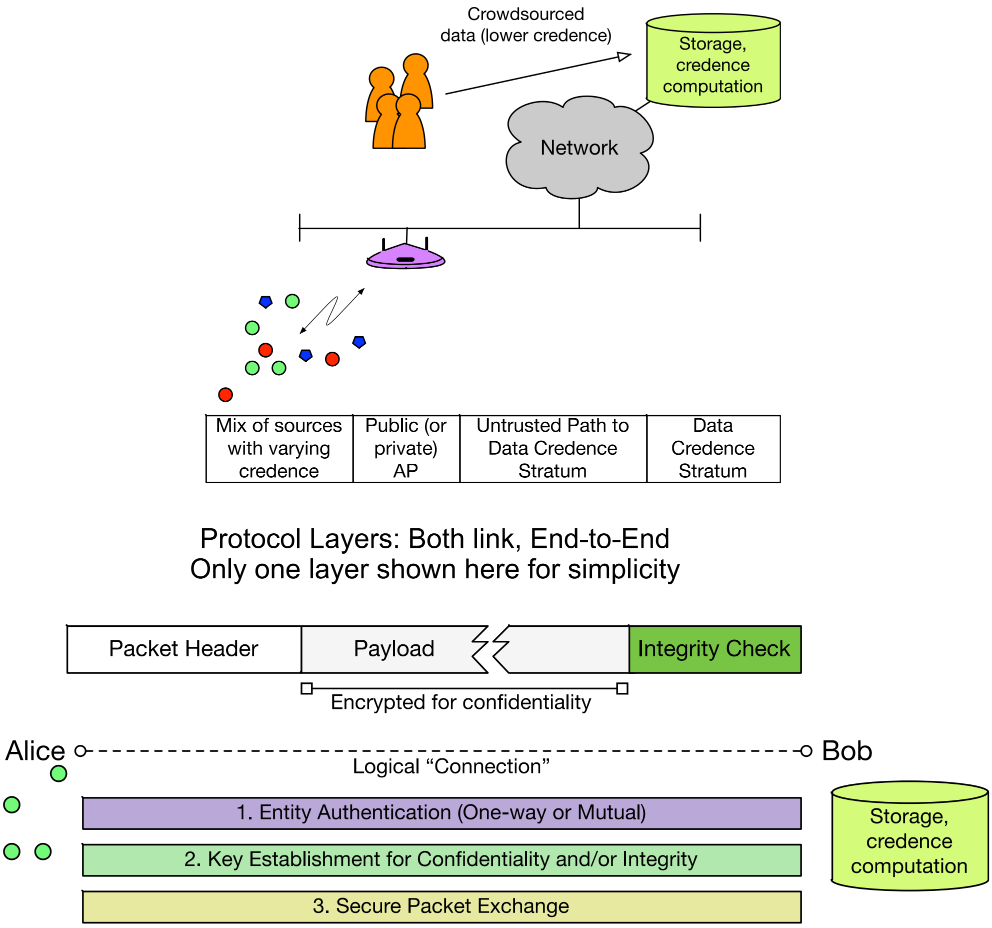


Figure 7: Credence in a mix of trusted and untrusted network paths

and conflict severity should be taken into account to evaluate the credence of each of the data sources.

We can evaluate the credence of each node in an Inconsistency Graph extending spreading activation models and Appleseed model [9] [46]. Under this approach we inject an initial energy to each node and have it propagated to other nodes along the IG edges until the energy distribution on all nodes converged. The IG edges are also updated continuously during the energy propagation based on the current energy distribution on each node as follows:

$$IG(i,j) = -log_{10}\left(\frac{Energy(j)}{\sum_{i=1}^{N} Energy(i)}\right)$$

Nodes with the higher final energy are considered less reliable and with lower credence.

We note here that the IG approach uses only the actual reports and not the subjective assessments of credibility (that may come from the <l,g> tuples). A challenge we envision is combining the graph approach with subjective logic, which may be possible with new methods for conjunction, consensus, and discounting proposed in [37].

4 TUNING DATA CREDENCE

In this section we consider major challenges in tuning the data credence stratum considered in previous sections. Challenges that exist include assessing credence in the presence of malicious actors, identifying relationships between data sources, and efficiency considerations. We elaborate on these challenges below.

Challenge 1: How do we authenticate for credence? The data credence stratum must consider the *authenticity* of the data using both network level security and cryptographic metrics. In the previous section, we introduced data credence metrics using subjective logic, and these have to be amended with security metrics in the presence of malicious actors. While there are no standard metrics for assessing the network level security, it may be possible to use quantifiable metrics, such as the number of hops (see for example [2]), link and end-to-end security protocols in use, key management in the system of interest (e.g., how fresh are keys) etc. An illustration of these issues is shown in Figure 7. The red sources have no authentication or integrity, the blue hexagons have only link-level authentication with a private access point (AP) and the green sensors have both link and end-to-end authentication.

Cryptographic metrics may also be characterized quantitatively based on the algorithms and key sizes (for example, using well known public estimates such as https://www.keylength.com). In the case of stored data, we suggest to include the effects of encrypted storage, storage ownership, and integrity checks of data blocks in models for credence assessment. These metrics may be incorporated in the internal credence model both locally (for data sources and blocks) and globally for a provider of different data sources. To the best of our knowledge, using both data and cryptographic credence at the same time has not been previously investigated.

Challenge 2: How do we assess relationships between data sources?
Often times, it is not clear as to which data source is related with which other data sources. Identifying the links for the *inconsistency graph* in Section 3 may not be a trivial problem. For example, sources that are close together in space should be part of the same group, but if the uncertainty in the location is very large, this may be an indication of whether or not it belongs to the set of data sources to be considered. We may know that a sensor is reporting soil moisture and it is in the vicinity of Fiona's farmland, but the precise location may not be available. We suggest to employ techniques from network science for this purpose. Research work in network science (see for example [10]) looks at group detection in social networks using stochastic models of link emissions from group entities and a maximum likelihood clustering.

Challenge 3: How do we improve data credence?
In some applications, it may be inevitable that deployment of additional sources in the field is the only option to improve data credence, because of the inconsistencies or sparsity of reports to adequately quantify the credence. For example, if the inconsistencies are due to large geographical separation, it may be useful to deploy additional sensors in the field to obtain higher granularity in space. In general, we expect the dimensions of the data credence assessment problem and the capability of the data sources to influence this additional deployment – as another example, if a sensor is unable to provide samples in time at a specific granularity, a duplicate sensor whose samples are offset in time may be an option.

Consider the example of the bank Bob and Fiona. If the data credence desired by the bank Bob is below an acceptable level, what strategies can Bob adopt? Two of the many possibilities are perhaps increasing the integrity of some of the less credible sources or adding more sources. For the sake of illustration, let us suppose that some of the sources deployed by Owen do not employ cryptographic protocols, raising the possibility that their reports could be modified reducing data credence. Addressing this may involve the use of integrity checks or better cryptographic techniques, or more granular samples in time – all of which may impact the battery life of sensors. Alternatively, only specific reports (either periodically or randomly) may be attached with integrity checks, which will enable the assessment of credence of the check-free reports. An alternative approach may involve either getting additional data from Oscar or Ogden or the deployment of additional sensors by the outsourced company Owen. The question that needs to be addressed is what strategies will provide the best result and are the most efficient in terms of deployment costs or energy costs at the sensors.

Challenge 4: How can we use cryptography efficiently for data credence?
As a second example consider the issue of auditing or forensics. Let us suppose that the storage cost of all of the collected data is unacceptable. How much data should be stored to have a specific level of data credence in the case of a needed audit? Should all of the stored data have integrity checks? Are there suitable data structures that can be used to reduce the storage/computational burden (see for example work that looks at *detecting* modifications in stored data in an untrusted cloud in [15])? These are open questions and challenges that need to be addressed for data credence.

5 TECHNOLOGIES ON THE HORIZON

In this section we describe two recent technological advances that introduce the potential for improving data credence but have inherent challenges as well.

Challenge 1: How can we exploit multiple link layer technologies for data credence?
There are multiple communications technologies that can support data exchange between "things". This includes long-range wireless RF-based technologies (such as GPRS, LoRa, Sigfox); short-range wireless RF-based technologies, free-space optical communications (e.g., visible light communications, Infrared-based communications), and wired

technologies (power line communications, Ethernet, optical fibers). Each link technology has its unique features and associated credibility (interface-dependent credibility). For example, data exchanged over wired medium is less likely to be spoofed compared to data exchanged over a wireless interface (assuming that it is hard to physically get access to the wired medium). Similarly, data exchanged through optical wireless interface is *confined in space*, and hence is inherently more secure than wireless radio-frequency interfaces [25] that may be attacked from farther distances.

Multiple communication interfaces can jointly be utilized in assessing and/or improving data credence. When multiple link technologies are exploited, the interface with higher reliability could help in enhancing the reliability of data exchanged over other interfaces. For example, visible light communications can be used to establish secure keys over RF links; consider a thing *A* that sends a master key to things *B* and *C* through visible light communications when they are in close proximity, then *B* and *C* use this key to establish session keys over wireless RF channels [39]. This enables secure key establishment and hence improves the reliability of data exchanged over RF.

Data exchanged over different interfaces can be correlated or fused (as in Figure 2, but even between the same pairs of things). By extracting correlated information and/or fusing data, the a priori credibility of information can be improved. For example, consider a network where some devices have an interface that allows communications using optical signals and another that uses RF. The advantage of using optical signals between things and a network infrastructure are: (a) ensuring that only things *within* the same physical location (where the information/data is relevant) will receive it; (b) minimizing interference on RF links; (c) enabling more reliable communications, since optical communications is less susceptible to attacks (e.g., eavesdropping - passive or replay- active attacks). In this case, under good weather conditions, the data exchanges over optical wireless channels may be associated with higher credence than data received over RF. In other words, the different *links* of a data source may have varying levels of internal credence!

The challenges here lie in adapting these technologies for improving data credence. Not all things are likely to have multiple interfaces, and there will likely be a mix of devices in a given environment/application. Efficiently deploying technologies to improve credence will be an ongoing challenge.

Challenge 2: What constraints/benefits do energy harvesting schemes bring to assessment of data credence?

The second emerging technology of interest is wireless energy harvesting. Devices can exploit the ever-increasing volume of wireless communications to harvest energy [26], hence prolong their lifetime. It is to be noted that different link technologies can be used for energy harvesting. In [35], energy is harvested from the received wireless optical signal, which is then used for transmitting RF signals. A simple view of energy harvesting is that devices have a “duty cycle” where things need time to recharge their batteries using ambient wireless signals which is significantly larger than the time for which they can transmit sensed data or take actions based on triggers. This duty cycle imposes constraints on the data credence (Is the sensed data sampled adequately?).

Among the challenges with energy harvesting are how things should be dispersed/deployed for satisfying a level of data credence for an application. For example, different “things” may be triggered with offset duty cycles, but such things may have varying levels of internal and external credence. Tuning and optimizing the deployment will have interesting problems to solve, in a manner similar to case of the multiple links.

6 Summary and Conclusions

We propose a vision with a systematic approach to maintain a data credence stratum assessing the credence or integrity of the data generated by Internet of Things. We suggest an integrated data credence analysis exploring data redundancy and data inconsistencies so as to provide automatic data credence assessment that also takes into account issues of information assurance (through cryptographic methods). We consider an approach to discover data inconsistency through the analysis of relationships between data items and data sources with additional metrics that may include cryptographic assurance. Our work is the first attempt to envision and utilize efficient inconsistency analysis and information assurance for implementing a scalable data credence stratum.

Acknowledgements

This work was partially supported by NSF BCS-1244672 grant.

REFERENCES

[1] S. Agarwal, A. M. Keller, G. Wiederhold, and K. Saraswat, “Flexible relation: An approach for integrating data from multiple, possibly inconsistent databases,” in *Proceedings of the Eleventh International Conference on Data Engineering*, 1995, pp. 495-504.

[2] A. Aksu, P. Krishnamurthy, D. Tipper, and O. Ercetin, “On security and reliability using cooperative transmissions in sensor networks,” *Mobile Networks and Applications,* vol. 17, no. 4, pp. 526-542, 2012.

[3] L. Bertossi, “Consistent query answering in databases,” *ACM Sigmod Record,* vol. 35, no. 2, pp. 68-76, 2006.

[4] L. E. Bertossi, and J. Chomicki, “Query Answering in Inconsistent Databases,” in *Logics for emerging applications of databases*, 2003, pp. 43-83.

[5] J. Bleiholder, and F. Naumann, “Data fusion,” *ACM Computing Surveys (CSUR),* vol. 41, no. 1, pp. 1, 2009.

[6] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi, “A cost-based model and effective heuristic for repairing constraints by value modification,” in *Proceedings of the ACM SIGMOD international conference on Management of data*, 2005, pp. 143-154.

[7] M. L. Brodie, “Data integration at scale: From relational data integration to information ecosystems,” in *Advanced Information Networking and Applications (AINA), 24th IEEE International Conference on*, 2010, pp. 2-3.

[8] F. Bry, "Query answering in information systems with integrity constraints," *Integrity and Internal Control in Information Systems*, pp. 113-130, 1997.

[9] M. Ceglowski, A. Coburn, and J. Cuadrado, “Semantic search of unstructured data using contextual network graphs,” *National Institute for Technology and Liberal Education*, vol. 10, 2003.

[10] G. B. Davis, and K. M. Carley, “Clearing the FOG: Fuzzy, overlapping groups for social networks,” *Social Networks,* vol. 30, no. 3, pp. 201-212, 2008.

[11] X. L. Dong, L. Berti-Equille, and D. Srivastava, “Integrating conflicting data: the role of source dependence,” *Proceedings of the VLDB Endowment,* vol. 2, no. 1, pp. 550-561, 2009.

[12] X. L. Dong, and F. Naumann, “Data fusion: resolving data conflicts for integration,” *Proceedings of the VLDB Endowment,* vol. 2, no. 2, pp. 1654-1655, 2009.

[13] X. L. Dong, B. Saha, and D. Srivastava, “Less is more: Selecting sources wisely for integration,” in *Proceedings of the VLDB Endowment*, 2012, pp. 37-48.

[14] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, “Corroborating information from disagreeing views,” in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 131-140.

[15] M. T. Goodrich, M. J. Atallah, and R. Tamassia, “Indexing information for data forensics,” in *ACNS*, 2005, pp. 206-221.

[16] J. Grant, “Classifications for inconsistent theories,” *Notre Dame Journal of Formal Logic,* vol. 19, no. 3, pp. 435-444, 1978.

[17] J. Grant, and A. Hunter, “Measuring consistency gain and information loss in stepwise inconsistency resolution,” *Symbolic and Quantitative Approaches to Reasoning with Uncertainty,* pp. 362-373, 2011.

[18] J. Grant, and A. Hunter, “Distance-based measures of inconsistency,” in *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pp. 230-241, 2013.

[19] L. Haas, “Beauty and the beast: The theory and practice of information integration,” in *International Conference on Database Theory*, pp. 28-43, 2007.

[20] C. O. Hernæs, "Banks should prepare for the Internet of Things," *TechCrunch*, https://techcrunch.com/2015/11/10/banks-should-prepare-for-the-internet-of-things/, accessed November 16, 2015.

[21] A. Hunter, and S. Konieczny, “Approaches to measuring inconsistent information,” *Inconsistency tolerance,* vol. 3300, pp. 191-236, 2005.

[22] A. Hunter, and S. Konieczny, “On the measure of conflicts: Shapley inconsistency values,” *Artificial Intelligence,* vol. 174, no. 14, pp. 1007-1026, 2010.

[23] T. Imieliński, and W. Lipski Jr, “Incomplete information in relational databases,” *Journal of the ACM (JACM),* vol. 31, no. 4, pp. 761-791, 1984.

[24] A. Jøsang, “A logic for uncertain probabilities,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 9, no. 03, pp. 279-311, 2001.

[25] A. Jovicic, J. Li, and T. Richardson, “Visible light communication: opportunities, challenges and the path to market,” *IEEE Communications Magazine,* vol. 51, no. 12, pp. 26-32, 2013.

[26] P. Kamalinejad, C. Mahapatra, Z. Sheng, S. Mirabbasi, V. C. Leung, and Y. L. Guan, “Wireless energy harvesting for the internet of things,” *IEEE Communications Magazine,* vol. 53, no. 6, pp. 102-108, 2015.

[27] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, “Truth finding on the deep web: Is the problem solved?,” in *Proceedings of the VLDB Endowment*, pp. 97-108, 2012.

[28] Z. Liu, H. A. Song, V. Zadorozhny, C. Faloutsos, and N. Sidiropoulos, “HFuse: Efficient fusion of aggregated historical Data,” in *Proceedings of SIAM International Conference on Data Mining*, 2017.

[29] C. Maag, "Officials reveal safety problems found at NJ transit before Hoboken crash," *USA Today*, accessed November 4, 2016.

[30] K. Mu, W. Liu, and Z. Jin, “A general framework for measuring inconsistency through minimal inconsistent sets,” *Knowledge and Information Systems,* vol. 27, no. 1, pp. 85-114, 2011.

[31] K. Mu, W. Liu, Z. Jin, and D. Bell, “A syntax-based approach to measuring the degree of inconsistency for belief bases,” *International Journal of Approximate Reasoning,* vol. 52, no. 7, pp. 978-999, 2011.

[32] J. Pasternack, and D. Roth, “Knowing what to believe (when you already know something),” in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 877-885, 2010.

[33] K. Pelechrinis, V. Zadorozhny, V. Kounev, V. Oleshchuk, M. Anwar, and Y. Lin, “Automatic evaluation of information provider reliability and expertise,” *World Wide Web,* vol. 18, no. 1, pp. 33-72, 2015.

[34] R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava, “Fusing data with correlations,” in *Proceedings of the ACM SIGMOD international conference on Management of data*, pp. 433-444, 2014.

[35] T. Rakia, H.-C. Yang, F. Gebali, and M.-S. Alouini, “Optimal design of dual-hop VLC/RF communication system with energy harvesting,” *IEEE Communications Letters,* vol. 20, no. 10, pp. 1979-1982, 2016.

[36] Y. Ren, V. I. Zadorozhny, V. A. Oleshchuk, and F. Y. Li, “A novel approach to trust management in unattended wireless sensor networks,” *IEEE Transactions on Mobile Computing,* vol. 13, no. 7, pp. 1409-1423, 2014.

[37] B. Škorić, S. J. de Hoogh, and N. Zannone, “Flow-based reputation with uncertainty: evidence-based subjective logic,” *International Journal of Information Security,* vol. 15, no. 4, pp. 381-402, 2016.

[38] S. Staworko, and J. Chomicki, “Consistent query answers in the presence of universal constraints,” *Information Systems,* vol. 35, no. 1, pp. 1-22, 2010.

[39] T. Sugiyama, "Information processing system, information processing apparatus and method, program, and recording medium," Google Patents, 2012.

[40] J. Wijsen, “Consistent query answering under primary keys: a characterization of tractable queries,” in *Proceedings of the 12th International Conference on Database Theory*, pp. 42-52, 2009.

[41] X. Yin, J. Han, and S. Y. Philip, “Truth discovery with multiple conflicting information providers on the web,” *IEEE Transactions on Knowledge and Data Engineering,* vol. 20, no. 6, pp. 796-808, 2008.

[42] X. Yin, and W. Tan, “Semi-supervised truth discovery,” in *Proceedings of the 20th international conference on World wide web*, pp. 217-226, 2011.

[43] V. Zadorozhny, and J. Grant, “A systematic approach to reliability assessment in integrated databases,” *Journal of Intelligent Information Systems,* vol. 46, no. 3, pp. 409-424, 2016.

[44] V. Zadorozhny, and Y.-F. Hsu, “Conflict-aware historical data fusion,” *Scalable Uncertainty Management,* pp. 331-345, 2011.

[45] V. Zadorozhny, P. Manning, D. J. Bain, and R. Mostern, “Collaborative for historical information and analysis: Vision and work plan,” *Journal of World-Historical Information: JWHI,* vol. 1, no. 1, pp. 1, 2013.

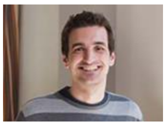
[46] C.-N. Ziegler, and G. Lausen, “Spreading activation models for trust propagation,” in *e-Technology, e-Commerce and e-Service, IEEE International Conference on*, Year, pp. 83-97, 2004.

AUTHOR BIOGRAPHIES

Vladimir Zadorozhny (www.sis.pitt.edu/~vladimir) is an Associate Professor of Information Sciences at the University of Pittsburgh School of Information Sciences. He received his Ph.D. in 1993 from the Institute for Problems of Informatics, Russian Academy of Sciences in Moscow. Before coming to USA he was a Principal Research Scientist in the Institute of System Programming, Russian Academy of Sciences. Since 1998 he worked as a Research Associate in the University of Maryland Institute for Advanced Computer Studies at College Park. He joined University of Pittsburgh in 2001. His research interests include information integration and fusion, complex adaptive systems and crowdsourcing, query optimization in resource-constrained distributed environments, and scalable architectures for wide-area environments with heterogeneous information servers. His research has been supported by NSF, EU and Norwegian Research Council. Vladimir is a recipient of Fulbright Scholarship for 2014-2015. He has received several best paper awards and has chaired and served on program committees of multiple Database and Distributed Computing Conferences and Workshops.

Prashant Krishnamurthy is an associate professor in the School of Information Sciences at the University of Pittsburgh. He has a Ph.D. in Electrical Engineering from Worcester Polytechnic Institute. He is a co-founder of the Laboratory of Education and Research in Security Assured Information Systems (LERSAIS), which has been designated as Center of Academic Excellence in information Education and Research (CAE + CAE-R) jointly by the United States NSA and DHS. His research interests include Wireless Network Security, Positioning and Localization, and Cryptography and Information Security. He has had research funding from the National Science Foundation, the National Institute of Standards and Technology, and The Army Research Office. He is the co-author of the textbook *Principles of Wireless Networks: A Unified Approach*.

Mai Abdelhakim is a Visiting Assistant Professor in the School of Computing and Information at the University of Pittsburgh. She received a Ph.D. degree in Electrical Engineering from Michigan State University in 2014, B.Sc. and M.Sc. degrees in Communications Engineering from Cairo University in 2006 and 2009, respectively. Prior to joining the University of Pittsburgh, she was a Postdoctoral Research Scientist at OSRAM Inc. (2015-2016). In 2006, she worked as an embedded software engineer in SySDSoft Inc. (currently Intel Mobile Communications). She also worked at the German University in Cairo (2007-2008) and at the Egyptian National Center for Radiation Research and Technology (2008-2010). Her research interests include cyber-physical systems security and communication networks design.



Konstantinos Pelechrinis received his PhD from the Computer Science department of University of California, Riverside, in 2010. His research interests include network science and he is involved in understanding the social, spatial, temporal and network dynamics of the behavior of people, developing models and algorithms for intelligent urban services as well as studying the effect of location-based social media on local businesses. He is a recipient of the Army Research Office Young Investigator (2015) for his work on composite networks.

Jiawei Xu is a PhD student of Information Science at the University of Pittsburgh School of Information Science. He received his Master Degree of Information Science with an emphasis in Big Data Analytics in 2015 from the same institute. His research interests include information integration, data fusion, and next generation databases.