RC25321 (WAT1210-082) October 25, 2012
Computer Science

IBM Research Report

Selecting Spatially-Relevant Information Providers

George Tychogiorgos
Electrical and Electronic Engineering
Imperial College
London SW7 2AZ
UK

Chatschik Bisdikian
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Selecting Spatially-Relevant Information Providers

George Tychogiorgos
Electrical and Electronic Engineering
Imperial College
London SW7 2AZ, UK
g.tychogiorgos@imperial.ac.uk

Chatschik Bisdikian
IBM Research
Thomas J. Watson Research Center
Hawthorne, NY 10514, USA
bisdik@us.ibm.com

***Abstract*—The on-demand binding between applications and information providers in loosely-coupled sensor-enabled systems raises the challenge for selecting the providers (i.e., sensor networks) supplying the most "relevant" sensory information. This paper focuses on spatial relevancy of sensory information determined by the quality and value of the desired and provided information. Specifically, the paper introduces a metric for spatial relevancy based on the concepts of quality of information (QoI) functions. We introduce expansion-proof descriptions of the QoI functions and we use those along with the the relevancy metric to (a) identify the most relevant provider among a collection of sensory information providers; and (b) select multiple providers with the objective to: ($b.1$) identify the minimum number of providers that cumulatively maximizes relevancy; and ($b.2$) considering the cost in engaging with providers, select the subset of providers that cumulatively maximizes the overall relevancy subject to a budgetary constraint. The performance and robustness of the proposed solutions are studied both analytically and by simulation for a number of provider topologies.**

I. Introduction

Consider the case where, say, a city agency needs to monitor air-quality (or, hazmat concentration levels, etc.) throughout the area of its authority. The agency would like to collect air-quality information at different quality levels, e.g., higher granularity in densely populated regions, and lower granularity at other regions. To collect the needed information, the agency uses sensors that it had deployed in the past. Unfortunately though, due to budgetary constraints and other logistics challenges, these sensors cover only portions of the area of interest. To supplement its information needs, the agency has decided to select and engage third-party fixed and mobile sensory information providers with whom it would create persistent or transient relations as necessary. The third-party providers could be other city agencies, private operators that, for example, monitor air-quality in public areas (parks, arenas, etc.), fleet operators whose fleet vehicles are equipped (for various reasons) with the necessary sensory devices, and even individuals whose smart-phones are capable of sensing air-quality conditions. This sensory information could be aggregated and "sold" to the city agency by sensory information brokers.

The above scenario exemplifies a trend where increased deployment and use of sensor networks is ushering a new era where information-rich solutions are becoming even more pervasive and integrated parts of our personal and professional lives. Applications such as environmental and habitat monitoring, infrastructure (highways, bridges, buildings, etc.) monitoring, security, surveillance and tracking, public transportation, traffic and utility management, commerce, manufacturing, food production, remote patient care and so on are just a few of the ever growing list of applications and market segments that are benefiting from the deployment of sensory infrastructure (and possibly contributed to our city-agency scenario). The emergence of the *Internet of Things* (IoT) [1] and *participatory sensing* [2] will further hasten the rate and ease with which information from tethered and untethered sensors, the Web, etc., will coalesce on demand to support our information needs via both loosely- and tightly-coupled sensor-enabled systems.

There are undoubtedly several challenges in realizing the "city agency" scenario. They relate, and not only, to technology; system (HW/SW) architecture and design, operation, and management; regulatory constraints; and, this being a city agency, public perception. It is the purpose of this paper to study one of these challenges that of dealing with selecting information providers that supply the most *relevant* information for the user's (e.g., the city agency's) needs. Specifically, we seek to establish procedures by which we can compare information sources based on how relevant the information they produce is to the desired and sought after information. To this end, we need to develop means to capture properties of the information against which relevancy can be assessed and metrics to capture the ensuing levels of relevancy.

Assuming semantically identical pieces of sought and provided information, [3] proposed using the spatiotemporal properties of information for identifying (or at least narrowing down) the relevant information. These properties also serve as the basis for *quality of information* (QoI) metadata [4] representing the physical context of the information, i.e., the time and space for which the information is applicable. In [3], relevancy was measured by "how spatiotemporally close" a piece of information provided was to the information desired. Specifically, spatial relevancy was measured by the degree

Research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

This paper was presented in part at the *12th IEEE Int'l Conf. on Mobile Data Management* (MDM), June 6–9, 2011, Lulea.

of overlap between the region R_p describing the coverage of sensory information from a provider and the region R_d describing the coverage of sensor information desired by a user.

As the number and variety of potential providers of information as well as the number of applications that depend on and search for them increases, the process of selecting the most relevant providers becomes more and more challenging. Furthermore, the fluidity of untethered sources (humans in participatory sensing, sensor-equipped vehicles, etc.) adds to the challenge as applications interested in information from a particular region may need to seek for and bind repeatedly to new(er) relevant sources. These challenges have a three-fold impact: increased processing, storage, and communication requirements, all of which raise concerns when considering resource-constrained sensor networks. The processing challenge is the obvious and direct one as more and more candidate sources have to be assessed and selected from. The other two are more subtle. The increase in the number of sources and applications will inadvertently result in an increase in the pertinent advertisements and exchange of metadata about (at least) the spatiotemporal and general QoI properties from the sources and/or desired by the applications. These metadata will also have to be stored at various nodes in the network.

There is an additional challenge that can further exacerbate all three previous challenges: *metadata expansion*. As more sources become available, new compound sources could (and would) be created as needed. For example, a new source reporting air-quality from the east (E) and north (N) regions of a city can be created by the combination of regional sources reporting air-quality from portions of the E, NE, and N regions of the city. How should the spatiotemporal properties of the compound source be represented? The obvious way is to combine (e.g., take the union of) the corresponding metadata from each of the constituent sources. This will result in a more populous entry for these metadata. As more and more sources are compounded this will lead to the unbounded increase of the related metadata entry, which of course will create major management burdens regarding their processing, communication, and storage.

Taking into consideration the aforementioned multitude of operational challenges, the contributions of our current work are:

- the introduction of QoI functions for describing the contextual desirability/quality of information;
- the definition of a novel problem and a new metric regarding information relevancy based on the QoI functions;
- the provision of finite, expansion-proof metadata descriptors for the QoI functions, using approximation techniques, such as spline surfaces;
- the formulation of optimization problems for selecting a single or multiple relevant providers with or without constraints; and
- the solution algorithms and study of these optimization problems.

The organization of the paper is as follows: Section II introduces relevancy and its QoI function based metric. Within the context of the relevancy definition and terminology introduced in Section II, Section III presents the scope of the problems to be addressed in the rest of the paper. Section IV presents the expansion-proof description of QoI functions. Section V introduces the multi-provider composition problem and studies pertinent optimization problems along with solution algorithms. Section VI provides the numerical evaluation of our solutions for various provider topologies. Section VII presents related work and we conclude with Section VIII with a summary of this work.

This paper extends [5], which focused on the multiple providers selection problem, by presenting in depth the relevancy metric (see Section II), including the definition of the problem scope (see Section III), and adding the expansion-proof description of QoI functions (see Section IV) along with the pertinent numerical results and prior work (see Sections VII and VIII.

II. The Relevancy of Sensory Information

We start with a brief summary of relevancy from [3], and then we built upon it focusing on a quality influenced definition of it. We then present the problems at hand, and discuss solution approaches in the following sections. For ease of presentation, and without lack of generality, we will study only two-dimensional regions; extensions to higher dimensionalityities are possible, albeit at increased levels of notational (and computational) complexity.

A. Background on Spatial Relevancy Metrics

In [3], we (implicitly) defined spatial relevancy as the degree of spatial overlap that there exists between the information sought and the information provided, e.g., the coverage of the sensor networks supplying the sensor data feeds that an application taps to. Consequently, we defined the metric r_s of *spatial relevancy* as:

$$r_s(R_p; R_d) = \frac{f\big(\mathcal{A}[R_p, R_d]\big)}{f\big(\mathcal{A}[R_d, R_d]\big)}; \tag{1}$$

where R_d is a description of the *desired* spatial properties of the information sought (the d-information) and R_p are the spatial properties for the information *provided* (the p-information); mnemonically, R could stand for *region*, but not necessarily.[1] $\mathcal{A}$ maps a correlation of the spatial properties of, say, R_d and R_p to the nonnegative reals (mnemonically, $\mathcal{A}$ could stand for the *area* of the overlapping regions R_d and R_p). Finally, $f(\cdot)$ is a non-negative, non-decreasing function of its argument, such as $f(x) = x$. The denominator $f\big(\mathcal{A}[R_d, R_d]\big)$ plays the role of a normalization coefficient so that $r_s \in [0, 1]$. Note that both the normalization coefficient and $f(\cdot)$ could have been incorporated in the definition of $\mathcal{A}$, however keeping them exposed aids the presentation.

[1] In [3], we wrote $r_s(R_d, R_p)$ instead. The slight change in notation in this paper is to emphasize the relevancy as a property of the p-information relative to the d-information.

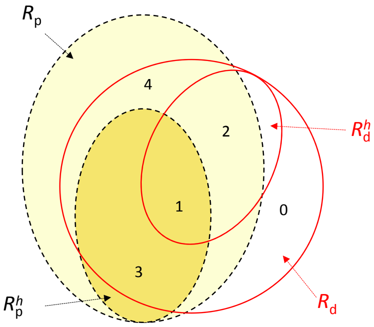


Fig. 1. Spatial properties of desired and provided sensor-originated information—regions are shown as ellipses for illustrative purposes only.

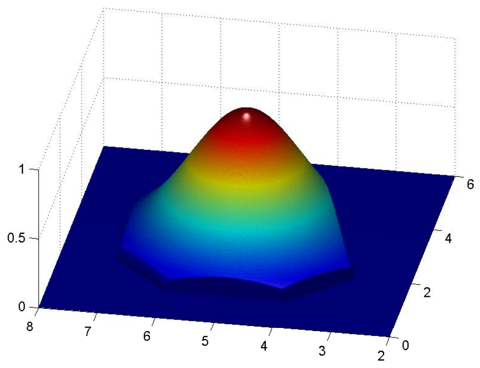


Fig. 2. An example of a desired (or provided) QoI function

As an example, consider Fig. 1 showing the spatial coverage R_d for the d-information in the regions enclosed by the solid red lines, and the corresponding region R_p for the sensor-generated p-information enclosed in the dashed black lines; ignore the superscript h for the moment. $\mathcal{A}[R_p, R_d]$ for this example represents the area of overlap between regions R_p and R_d, and assuming $f(x) = x$:

$$r_s(R_p; R_d) = \frac{area[R_p \cap R_d]}{area[R_d]}. \tag{2}$$

B. Generalizing Spatial Relevancy

Moving beyond [3], suppose that there is a desire for *higher* accuracy in the information from the subregion R_d^h of region R_d, e.g., receiving images of higher resolution, or detecting events of interest with higher probability, etc. Likewise, suppose that sensor feeds have two levels of accuracy, *high* level at the subregion R_p^h of region R_p and *low* elsewhere. Then, we can extend (2) and write:

$$r_s(R_p; R_d) = \frac{\sum_{i\in\{h,l\}} \sum_{j\in\{h,l\}} \alpha_i \cdot \beta_j \cdot area[R_d^i \cap R_p^j]}{\sum_{i\in\{h,l\}} \alpha_i \cdot area[R_d^i]}, \tag{3}$$

where $R_k^l = R_k \setminus R_k^h$, $k \in \{d, p\}$. The α's and the β's are relative weights describing the level of desirability and/or utility for data received from the corresponding areas. These weights can further be selected to normalize the range of $r_s(R_p; R_d)$ in $[0, 1]$, where $r_s(R_p; R_d) = 1$ represents perfectly relevant information, i.e., the information provided was of quality equal (or better) than desired across the entire region R_d. The four regions $R_d^i \cap R_p^j$ correspond to the four regions R_1–R_4 shown in Fig. 1 and together with the region R_0 they form a partition of the desired region R_d; $R_0 = R_d \setminus R_p$ and by convention its relative weight is 0.

It should be clear that one can keep on extending (3) by adding new gradations of desired and provided information qualities and associate them with corresponding intersect regions $R_d^i \cap R_p^j$ and weight products $\alpha_i \cdot \beta_j$ (or, in general, through weight functions $w(\alpha_i, \beta_j)$). However, in generalizing the above, we take a slightly different approach, while still keeping the region "overlap" principle of this and previous subsections.

C. The QoI Functions

By adding gradations in the desirability or the QoI across the regions R_e, $e \in \{d, p\}$, we can generalize the "overlap" principle in (2) and, consequently, the information spatial relevancy definition and metric.

Specifically, let $\omega = (x, y)$ be a point in a two dimensional region R and let

$$q_d : \omega \in R_d \to [0, 1], \quad \text{with} \int_{R_d} q_d(\omega)\, d\omega < \infty \tag{4}$$

be a *desired QoI function* (or a QoI d-function) describing the QoI of the desired d-information at point ω. For example, at point ω_0, a detection application "desires" to receive information about occurrences of events of interest where the probability of correctly detecting an event be z_0 (i.e., $q_d(\omega_0) = z_0$), or, the concentration of air-pollutants at point ω_0 shall has an error no larger than ϵ_0. The range of q_d could be the entire real line, but we assume that it is expressed in relative terms and is normalized to $[0, 1]$ with values closer to 1 representing higher quality levels for the d-information. By convention, we set $q_d(\omega) = 0$, for all points ω outside the desired region R_d.

Fig. 2 shows an example of a QoI d-function, where an application is highly interested in a concentrated area (i.e., requires information of high accuracy for that area). The interest tempers off away from that area and eventually drops to zero (outside the R_d region).

We can define the *provided* (or *provider*) *QoI function* (the QoI p-function) $q_p(\omega)$ on a region R_p in a completely analogous fashion to $q_d(\cdot)$ and R_d. Fig. 2 applies in this case as well. For example, the figure could represent event detection capability of a sensor network comprising of one or more sensors concentrated in a particular region, hence, having high detection accuracy in that area and decreasing accuracy away from the area. In the sequel, when we do not need to distinguish between the QoI d- and p-functions we will collectively refer to them as the "QoI functions" $q(\cdot)$ dropping the indexes d and p from the pertinent notation.

According to one operational mode, sensor-enabled applications may "announce" their information needs by broadcasting their QoI d-function $q_d(\cdot)$ and its support region R_d; interested providers may then respond to the application in kind. According to another operational mode, providers may "advertise" their sensing capabilities by broadcasting their QoI p-function $q_p(\cdot)$ and its support region R_p; applications may then sift through these advertisements and select appropriate providers.

These (or other) operational modes are beyond the scope of this paper. We are concerned only with the fact that an application ends up with a collection of QoI p-functions from various providers. Based on them and its own QoI d-function, it assesses their relevancy to its information needs and chooses an appropriate one (or ones) based on some operational criteria.

D. The QoI Function-based Relevancy Metric

We are now ready to extend the relevancy metrics of previous subsections as follows:

$$r_s^v(q_p; q_d) = \frac{\int_{R_p \cap R_d} v\big(q_p(\omega); q_d\big)\, d\omega}{\int_{R_d} v_d\big(q_d(\omega); q_d\big)\, d\omega}, \tag{5}$$

where $v(\cdot; q_d)$ is a (non-negative) *value* function that represents the value the sensor-enabled application gains in executing its task when it uses information of quality $q_p(\omega)$ at point ω. The notation $v(\cdot; q_d)$ implies that, in general, the value function can be expressed relative to q_d, as was done, in a different context, with the *QoI satisfaction index* in [6]. The denominator in (5), which is assumed to be finite, plays the role of a normalization factor so that $r_s^v(q_p; q_d) \in [0, 1]$. We silently assume that an application gains nothing extra if it receives information of higher quality than what it asked for, and, hence, for each $\omega \in R_d \cap R_p$: $v\big(q_p(\omega); q_d\big) \in [0, v_d\big(q_d(\omega); q_d\big)]$. If the latter is not the case, one may need to appropriately redefine the normalization role of the denominator; we do not consider the latter case in this paper.

To reflect intuition, the value function is (selected) such that the relevancy metric exhibits an increasing trend with q_p. Specifically, if there are two providers p_1 and p_2 with p_1 "closer" to the desired needs of the application than p_2, i.e., their QoI functions satisfy:

$$\|[q_d - q_{p_1}]^+\| \leq \|[q_d - q_{p_2}]^+\|, \text{ where } x^+ \overset{\text{def}}{=} \max(0, x), \tag{6}$$

according to some function norm operator defined over the set R_d, e.g., the l_2 norm, then:

$$\int_{R_{p_1} \cap R_d} v\big(q_{p_1}(\omega); q_d\big)\, d\omega \geq \int_{R_{p_2} \cap R_d} v\big(q_{p_2}(\omega); q_d\big)\, d\omega. \tag{7}$$

For the numerical results later in the paper, we will use "$\min\{\cdot\}$" as the value function. In this case, we write (we drop the superscript v for brevity):

$$r_s(q_p; q_d) = \frac{\int_{R_p \cap R_d} \min\big\{q_p(\omega), q_d(\omega)\big\}\, d\omega}{\int_{R_d} q_d(\omega) d\omega}. \tag{8}$$

The numerator in the above expression can be thought as the "volume" of the provider capability level function measured only for the part of R_d that the provider can support, which is the intersection of the areas R_d and R_p. In addition, the $\min\{\cdot\}$ term is used in order to make sure that values of provided accuracy higher than what is desired will not be taken into consideration in spatial relevancy calculations. The denominator is a normalization factor so that $r_s \in [0, 1]$.

In closing this section, we remark that (5) can be interpreted as representing the aggregate, normalized utility that can be achieved in relation to d-information when receiving p-information from provider p. There might be alternative interpretations of (5), such as probability expectations of some form, or the conditional or relative entropy of the d-information in the presence of the p-information [7]. There could be some operational complications that these interpretations may introduce, such as the need for a priori knowledge or on-demand computation of joint or conditional probability densities between entities (the providers and the applications) that had no prior kinship to each other. Nonetheless, in principle, these alternative interpretations do not alter the fundamentals of advertising desired or provided QoI functions and making provider selections based on them, which is the premise of this paper.

III. Problem Scope Definition

Ideally, communicating and manipulating general functions such as q_d and q_p defined over general sets R_p and R_d in order to calculate the relevancy metric in (5) could require unpredictable (if not infinite) accuracy, storage, and computational resources. Operationally, dealing with infinite accuracy is impossible, and dealing with unpredictable quantities is highly undesirable. Practically, the QoI functions and corresponding support regions R_e, $e \in \{d, p\}$, will be characterized approximately and communicated by a collection of (approximation) descriptors. We will refer to these descriptors as the *QoI coverage metadata*.

Clearly, approximate geospatial descriptions of regions, based on various types of polygon representations, can serve the above purpose for the boundaries of the R_e regions [8]. The geospatial descriptions are typically aid in answering queries regarding topological relationships such as when querying whether a point ω or a region $\mathcal{A}$ is internal, external, at the boundary, or intersecting another region $\mathcal{B}$. Owed to the fact that we are dealing with region intersections as well, such as $R_p \cap R_d$, topological relationships play role in our case as well. However, our queries are not topological in nature, at least not in the traditional sense. Our objective is to order and select providers based on relevancy assessed over the intersection of support sets for the QoI functions q_e, $e \in \{p, d\}$, and of course their values, see (5), all described by the QoI coverage metadata. These metadata will be communicated and stored at recipient nodes, e.g., the application node, or a provider registry, or other intermediary nodes.

Due to the nature of the situation, the QoI p-functions (and their support sets) could be the result of aggregation of constituent QoI p-functions and regions. For example, a "super-provider" may be formed to represent the accumulation of contributing providers in a set $\mathcal{P} = \{p_1, p_2, \ldots, p_{|\mathcal{P}|}\}$.

Thus, for example, if the QoI coverage metadata entry in the registry of a provider accommodates only up to M elements, then, this number is bound to be exceeded if the QoI coverage metadata for the super-provider are simply the union of the metadata of the constituent providers. Therefore, owed to the latter fact, a predictable structure for these metadata will also be required.

Hence the information relevancy problem at hand is summarized as follows:

- *Summarize the QoI functions of providers and applications through finite-sized, expansion-proof descriptors (metadata).* Using these descriptions,
- *Assess the relevancy of providers to an application's needs.* Using these assessments,
- *Select one or more of the providers to satisfy the application's needs given selection criteria, such as the most relevant provider, or the most relevant collection of providers given a budget (e.g., energy, cost) constraint.*

In the next section, we consider the expansion-proof descriptors which along with their evaluation presented in the numerical results Section VI covers the sub-problem of selecting the single most relevant provider. The multi-provider selection sub-problem will be covered in Section V.

IV. Expansion-proof QoI Function Description

As stated before, due to the generality of QoI functions, their communication, storage, and processing requirements may be quite unpredictable which has severe implication in managing system resources effectively. Hence, it would be desirable to describe them in a way that ensures predictable utilization of system resources while acknowledging their role in the process of selecting the most relevant information providers to serve an application's needs based on the relevancy metric r_s^v in (5).

To this end, we present two ways to describe QoI functions using a collection $\mathcal{M}$ of QoI coverage metadata of finite size $|\mathcal{M}| = M$. One way is based on describing the QoI functions discretely by sampling them, while the other involves interpolating between samples using splines. Note that the QoI coverage metadata (which, for brevity, will be referred to simply as metadata) could be part of the bigger collection of QoI metadata in [4]. The size M is a design parameter seeking to balance efficiency and accuracy in describing quality functions.

The two methods discussed next are clearly not the only ones. Other approaches were considered (such as descriptors based on contours) but not discussed here. The methods discussed offer intuitive simplicity (sampling) and established flexibility and utility (splines). Performance comparisons between these techniques are included in Section VI.

A. Sampling-based QoI Function Description

Sampling is a widely used method to approximate continuous functions while controlling the amount of generated data [9]. Sampling has found numerous applications in signal processing over the last years where a continuous signal must be stored and processed by a finite accuracy computer system [10]. *Uniform sampling*, where samples are uniformly distributed, is used for smooth functions while *non-uniform sampling* is often preferred to approximate functions with abrupt changes in subsets of their range. Given the sampled values, the original continuous signal can be approximated by using interpolation techniques such as *linear*, *nearest neighbor*, and *cubic* interpolation [11]. For the numerical evaluations later on, we use uniform sampling along with linear interpolation, an example of which is shown in Fig. 3 for the simple case of an one-dimensional QoI function defined over a line segment.

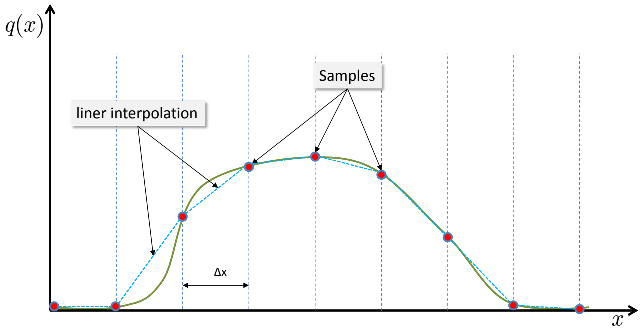


Fig. 3. Uniform sampling with linear interpolation for a 1-d QoI function.

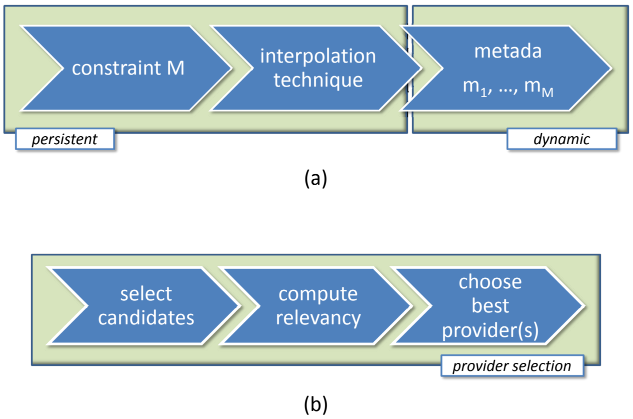
Sampling can be used to approximate the QoI functions q with a maximum of M parameters. Since we have defined the QoI functions on $\mathbb{R}^2$, the M parameters can be generated by sampling the continuous q function at $M/3$ points (either uniformly or non-uniformly) and storing a triplet $(x, y, q(x, y))$ for each sample; x and y are used to store the coordinates of the sample and $q(x, y)$ is the (desired or provided) quality value at the sampling point. As expected, the larger the number of sampling points are the more accurate the approximation will be. These M parameters can be communicated so that other parties can use an interpolation method to generate the approximation points $\hat{q}$.

The next subsection presents an alternative approximation method based on *splines*. The way that these two methods can be used within the "city agency" scenario and the comparison of their approximation performance will also be also explained in the sections to follow.

B. Spline-based QoI Function Description

Splines are piecewise polynomial curves which are differential up to a prescribed order [12]. A *B-spline* has the property that every spline of a given polynomial degree can be expressed as linear combination of a set of B-splines of the same degree. The *B-spline surfaces* are the result of the tensor product of B-spline curves, where a tensor product surface is generated by:

$$p(x, y) = \sum_{i=1}^{K} \sum_{j=1}^{L} \alpha_{ij} B_i(x) B_j(y), \tag{9}$$



(a)

(b)

Fig. 4. (a) Exchanged information regarding QoI function; and (b) selection of providers.

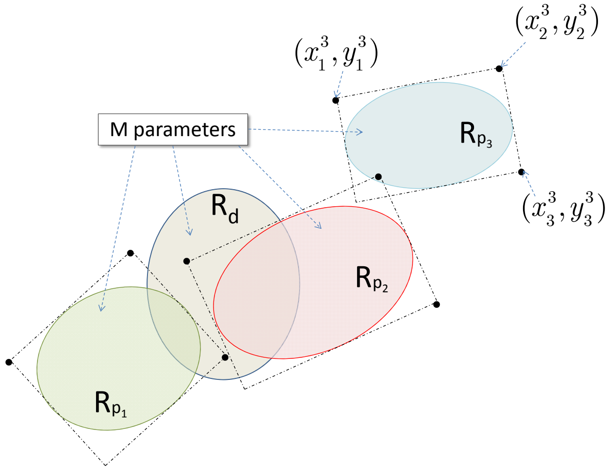


Fig. 5. Example of multiple desired/provided regions R and containing rectangles.

with $B_i(\cdot)$ and $B_j(\cdot)$ independent spline curves that form a basis and α_{ij} the spline control points.

The construction of the B-spline curves $B_i(\cdot)$ and $B_j(\cdot)$ is a two-pass process (one for each variable) and is based on the calculation of the so-called *knot* vectors and *control points* α_{ij} [13]. The design parameters of the method are the size of the knot vectors, n^x_{knot} and n^y_{knot}, and the spline order along the x and y dimensions, $order_x$ and $order_y$. The spline order is in essence the order of the polynomial used for the approximation. The input of the approximation procedure is the sample matrix of the QoI function, $\tilde{q}$, along with the sampling vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.[2] The resulting finite description of q consists of M parameters, the knot vectors of size n^x_{knot} and n^y_{knot}, and a matrix of size $(n^x_{knot} - order_x) \times (n^y_{knot} - order_y)$ containing the *control points*, α_{ij}. Thus, the finite description of q will be $P = n^x_{knot} + n^y_{knot} + (n^x_{knot} - order_x) \cdot (n^y_{knot} - order_y)$ points. These M parameters are necessary to be communicated so that the other parties can generate the approximated sample points, $\hat{q}$.

Due to their smooth, differentiable behavior, and ease of construction, splines and spline surfaces have long been studied and are popular in approximating single- and multivariate functions. Because they can be described by a finite number of points, they are also our preferred approximation choice in describing QoI functions q. Since, we define q on $\mathbb{R}^2$, we make use of spline surfaces as in (9) and use the aforementioned M control points and knots to describe it. Increasing the number of *knots* of the approximation or the order of the *spline*, and, hence, the number of *control points*, and eventually M, would give better approximation results. However, the simulation results presented later on show the efficiency of the method even for low order approximations.

C. Advertising the QoI Functions

While the specifics of how the providers and the consumers, e.g., the city agency of our motivating use case, exchange their QoI function descriptions is outside the scope of this paper, we briefly discuss here what information is to be exchanged. Fig. 4.a summarizes the information. This shared information is grouped in two categories: persistent and dynamic information. The *persistent* information represents information that changes infrequently (if at all), such as the constraint on the size M of metadata to be exchanged as well as the interpolation technique to be used. Persistent information can be thought as part of system-wide information that could be configurable and does not need to be exchanged frequently, e.g., the city agency may announce that it deals only with spline interpolation for QoI function descriptions of size M. The *dynamic* information pertains to the QoI coverage metadata that providers and possibly consumers may need to exchange on demand.

Next we summarize the steps followed in selecting providers, highlighted in Fig. 4.b in conjunction of the advertised information noted in Fig. 4.a:

- Providers encode their QoI p-functions $q_p(\cdot)$ using an agreed upon approximation technique, such as sampling or spline surfaces, based on M parameters. Being good citizens, they also (optionally) calculate the minimum rectangular $\mathfrak{R}(R_p)$ containing the provided region R_p. This requires three additional (x, y) points. The optional rectangle $\mathfrak{R}(R_p)$, which requires a total of 6 additional pieces of metadata, can be used to quickly narrow down the candidate providers, see shortly. A consumer may encode its own QoI d-function $q_d(\cdot)$ likewise.
  In reference to Fig. 5, the selected approximation method is used to generate M parameters describing $q_e(\cdot)$ in region R_e, $e \in \{d, p\}$. Three additional points $\{(x_i, y_i); i =$

[2] The QoI function q, the sample matrix $\tilde{q}$ and the sampling vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are connected by: $\tilde{q}\,(i, j) = q\,(\boldsymbol{x}\,(i)\,, \boldsymbol{y}\,(j))$.

$1, 2, 3\}$ are also used (optionally) to describe the minimum rectangle containing these regions.

- A consumer may distribute/advertise a CfP (*Call for Providers*) along with its M (or, $M + 6$), parameters of its own QoI d-function and collects responses from providers; alternatively, a consumer may passively listen to providers distributing/advertising their QoI coverage metadata.
- (optionally) The consumer may quickly filter out any provider p whose minimum rectangle $\Re(R_p)$ does not intersect with its own minimum rectangle $\Re(R_d)$—note that topological operations involving rectangles are very straightforward and easy to implement.
- The consumer uses the M parameters it receives to approximate the QoI p-function of any remaining provider (the "candidates" noted in Fig. 4.b) using the agreed upon approximation technique, e.g., interpolating between sampled points or by generating $B_i(\cdot)$ and $B_j(\cdot)$, based on the knot vectors for variables x and y, respectively, and (9). Finally,
- The consumer determines each provider's relevancy by using the approximated QoI functions in (5); providers may then be ranked accordingly.

The above procedure assumes that consumers and providers interact with each other directly. However, it is quite possible that a consumer, e.g., the city agency, may delegate the selection process to a proxy, or even a collection of proxies, e.g., city sub-agencies, acting on its behalf. Likewise, a provider may be a logical entity representing (or, brokering for) a collection of actual sensory information providers. In these cases, the selection of relevant providers is accomplished at the granularity of the QoI function exchanged and what entities these represent.

In closing this section, we note, that whether:

- the CfP contains just the M parameters, or just the 6 rectangle parameters, or all $M+6$ parameters; or whether
- a provider pre-calculates its q_p approximation, post-calculates it based on the CfP, e.g., use its M points to describe q_p only in the region of interest and not on the entire R_p; or whether
- the agency and the providers communicate with each other directly or through a proxy/broker in the middle, as note earlier; or whether
- only providers need to encode their q_p; or whether
- any other related design choices are considered

are outside the scope of the current paper and left for future investigations. Here we focus only on the fundamental structures and procedures of the relevancy assessment and provider selection on top of which all the other choices can be considered and evaluated.

As earlier noted, Section VI contains performance comparisons of the two QoI function approximations within the context of selecting the most relevant provider. The next section develops the framework for selecting instead groups of providers that, in aggregation, are the most relevant one based on selection criteria.

V. Multi-Provider Consideration

While it is possible that a single provider may suffice in satisfying an application's needs, it is quite likely that it will not. In this case, it would be desirable to be able to judiciously select a number of providers that cumulatively provide the most relevant information.

Using our finite-size, expansion-proof metadata principle, in this section, we consider the composition of sensory information providers and the selection of the most appropriate set of providers based on criteria such as maximum coverage and maximum aggregate geospatial relevancy for a given constraint. In the context of the city agency scenario, this may correspond to the case that the city agency will have to select the most appropriate providers given a budget constraint.

In general, we assume an application with q_d and R_d representing its desired QoI d-function and corresponding region. There is also a set $\mathcal{P}$ of providers of size $|\mathcal{P}| = N$ with $q_i(\cdot) \overset{\text{def}}{=} q_{p_i}(\cdot)$ and $R_i \overset{\text{def}}{=} R_{p_i}$, $i \in \{1, \ldots, N\}$, representing the corresponding provider QoI p-functions and regions. Note that, whenever the context permits it, in this section we will drop the p from provider-related entities for notation brevity and, instead, the index i will represent provider p_i.

In the following subsections, we consider two cases: (*a*) the no-cost case, where we seek to find the minimum number of providers that satisfy the application needs without any budgetary constraints; (*b*) the cost case, where engaging providers comes at a cost and applications (for example, the aforementioned city agency) have budgetary constraints. In both cases, we will first formulate a model for the problem and then consider a solution for it.

A. Maximum Relevancy with Minimum Providers and No-cost

We start with the case of selecting the minimum number of providers that can cover as much of the desired region as possible while attaining as high QoI as possible. To this end, let $\mathbf{I} = [I(1), \ldots, I(N)]$ be the provider *selection indicator vector*, where

$$I(i) = \begin{cases} 1, & \text{if provider } i \text{ is selected;} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Additionally, let the aggregate provider region $R_p^{\mathbf{I}}$ be the union of all the selected provider regions, i.e.,

$$R_p^{\mathbf{I}} = \bigcup_{i=1}^{N} I(i) \cdot R_i, \tag{11}$$

and let $R_p \overset{\text{def}}{=} R_p^{\{\mathbf{I}=\mathbf{1}\}} = \bigcup_{i=1}^{N} R_i$.

The selection of the appropriate set of providers to maximize the coverage of the desired region with no cost can be modeled by the following optimization problem Π_0:

Problem $\mathbf{\Pi_0}$: For $I(i) \in \{0,1\}$, $i \in \{1,\ldots,N\}$,

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{N} I(i), \quad \text{such that, } \forall \omega \in R_d \cap R_p : \\ &(1) \quad \sum_{i:\omega \in R_d \cap R_i} I(i) \geq 1; \text{ and} \\ &(2) \quad \max_{i:\omega \in R_d \cap R_i} \big[I(i) \cdot q_i(\omega) \big] = \max_{i:\omega \in R_d \cap R_i} \big[q_i(\omega) \big]. \end{aligned} \tag{12}$$

Constraint (1) is a *coverage* constraint that states that for each point $\omega \in R_d$ covered by one or more providers, at least one of them will be selected. Constraint (2) is a *preference* constraint that states that the provider with the highest QoI at a point ω shall be chosen. Note that this model allows the selection of providers that overlap at some points, however, it assures that the best provider at each point is among the selected ones. Therefore, the formulation is implicitly maximizing the aggregated spatial relevancy.

Problem Π_0 is a generalization of the *set covering* problem [14] on three dimensions (each 2D point ω is also associated with QoI value $q_d(\omega)$) and for unity costs. The set covering problem relates to finding the minimum number of sets whose union includes all points of the "universe." It is modeled by the following *integer programming* formulation: For $I(s) \in \{0,1\}$, for all $s \in \mathcal{S}$,

$$\text{minimize } \sum_{s \in \mathcal{S}} c(s) \cdot I(s), \text{ such that } \sum_{s:e \in s} I(s) \geq 1, \tag{13}$$

for all elements $e \in \mathfrak{U}$, where $\mathfrak{U}$ is the universe of points, $\mathcal{S}$ is a family of subsets of $\mathfrak{U}$ and $c(s)$ is the cost associated with set s in $\mathcal{S}$. The *set covering* problem is one of Karp's 21 *NP*-complete problems [15]. Therefore, the Π_0 problem is *NP*-complete as well and, hence, there is no polynomial-time algorithm that solves it. The most efficient algorithm solving (approximately) the set covering problem is a *greedy* algorithm that is based on the following simple operation: At every iteration, choose the set that contains the largest number of uncovered elements. The algorithm terminates when all elements are contained in the sets already selected.

Based on the aforementioned iterative operation, we propose a solution to problem Π_0 described by Algorithm 1 which, at each iteration, selects the most appropriate subset of providers that maximize the total relevancy with respect to the desired information, which is described by the QoI function q_d. Because of the possibility that $R_d \cap R_p^{\mathbf{I}}$ contains infinitely many points, the selection criterion at each iteration is not the number of points contained in each set of providers but, instead, the increase in the spatial relevancy metric. Thus, the provider that results in the largest increase in the aggregate relevancy is chosen at each iteration and the algorithm terminates when none of the remaining providers can increase the aggregate relevancy further.

More specifically, at each iteration t, the aggregate region $\mathcal{S}$ of the already selected providers $\mathcal{F}$, i.e., $\mathcal{S} = \cup_{k \in \mathcal{F}} R_k$, is merged with the new candidate region R_i. Then, the relevancy of the aggregated QoI p-function $q_i^{\mathcal{F}}(\omega)$ (explained shortly) is calculated for each candidate provider i when aggregated with the providers in the set $\mathcal{F}$ of already selected ones. Consequently, the provider leading to the highest aggregate relevancy, V^t, is selected, until there is no further increase in the total relevancy.

Algorithm 1 – Aggregate Relevancy

```
1: Initialize: $\mathcal{F} = \emptyset$, $\mathcal{S} = \emptyset$, $\mathcal{P} = \{1,\ldots,N\}$, $t = 1$ and $V^0 = 0$;
2: Set: $\mathcal{F}_i^t = \mathcal{F} \cup \{i\}$, $\mathcal{S}_i^t = \mathcal{S} \cup R_p^i$ for all providers $i \in \mathcal{P}$;
3: Calculate spatial relevancy, $r_s^t(i) \overset{\text{def}}{=} r_s^t\big(q_i^{\mathcal{F}}(\omega); q_d(\omega)\big)$, for all regions $\mathcal{S}_i^t$ using equation (5);
4: Set: $k \leftarrow \arg\max_i \{r_s^t(i)\}$ and $V^t \leftarrow r_s^t(k)$;
5: if $V^t = V^{t-1}$ then
6:     STOP;
7: else
8:     Set: $\mathcal{F} \leftarrow \mathcal{F}_k^t$, $\mathcal{S} \leftarrow \mathcal{S}_k^t$; $\mathcal{P} \leftarrow \mathcal{P} \setminus \{k\}$;
9:     Set: $t \leftarrow t + 1$;
10:    GOTO step 2;
11: end if
```

In step 3 of Algorithm 1, we use the *aggregated* QoI p-function $q_i^{\mathcal{F}}(\omega)$ which represents the *collective* QoI behavior of the already selected providers (in the set $\mathcal{F}$) and the new candidate provider i at the point $\omega \in \mathcal{S}$. Specifically, given two providers i and j with $q_i(\cdot)$, $q_j(\cdot)$, R_i and R_j their respective QoI p-functions and coverage regions, their aggregated QoI p-function $q_i^j(\cdot)$ is defined on $R_i \cup R_j$ where $q_i^j(\omega) = h\big(q_i(\omega), q_j(\omega)\big)$; recall that a QoI p-function is set to 0 outside its region R_p. The transformation h produces another QoI function from the constituent QoI functions which reflects how the quality of fused information is assessed. For example, if the accuracy of a measurement from provider i at a point ω is 97% and from provider j is 95%, the aggregated QoI from the two providers could be taken to be the best of the two, i.e., 95%, hence, "$h \equiv \max\{\cdot\}$." We use the latter example h in our numerical results later on, thus for $\omega \in R_i \cup R_j$, we will use:

$$q_i^j(\omega) \overset{\text{def}}{=} h\big(q_i(\omega), q_j(\omega)\big) = \max\big\{q_i(\omega), q_j(\omega)\big\}. \tag{14}$$

It should be noted here that it is cases like the above where providers may aggregate their information provision services that led to the consideration of managing the number of parameters used to describe the QoI functions and corresponding coverage regions. If these parameters were left to simply accumulate over successive (and unpredictable) provider aggregations would have resulted in the unpredicted and unbounded number of parameters mentioned in the previous section. Specifically, with regard to the QoI descriptions in the aforementioned provider aggregation, we assume that the values of $q_i(\omega)$ and $q_j(\omega)$ are reconstructed approximately from their finite-sized parameter representations, e.g., the spline surfaces, before calculating $q_i^j(\omega)$. Should the resulting aggregate QoI p-function needs to be stored or communicated, it will be done so using its own bounded-sized QoI description representation.

Algorithm 1 can be implemented in polynomial time. At each iteration, the algorithm determines the optimal provider to select, but, similarly to how the *greedy* algorithm behaves for the original set covering problem, this may not always lead to the overall optimal solution. However, it is easy to prove that the performance of the algorithm is upper bounded by a function of the optimal solution and the number of points in the universe $\mathfrak{U}$ [16]. In fact, this bound is $m \ln(n)$, where m is the optimal number of providers and n is the number of sample points trying to cover.

The problem model described in this section did not take into consideration the possible cost for engaging with information providers. Problem formulation Π_0 and its solution in Algorithm 1 identify the best subset of providers that maximize the aggregate spatial relevancy of information independently of the cost. Next we consider a problem model where cost plays role in choosing the optimal provider set.

B. Maximum Relevancy with Budget Constraints

Since nothing comes for free, the consumer, and certainly the city agency, will have to face the realities of budgetary constraints sooner or later. In this case, we assume that the consumer has a finite budget reserve B and engaging with provider i costs c_i, $i = 1, \ldots, N$. The cost and the budget could be in the form of monetary cost, e.g., a fee paid to use the services of a provider, or resource cost, e.g., energy consumed when engaging with a provider. Furthermore, the cost c_i could be a flat rate that the provider charges or a contracted price reflective of the attained relevancy $r_s(q_i; q_d)$; we will not delve further on this issue.

Given such a budgetary constraint, we are now interested in finding the optimal set of providers that will maximize the spatial relevancy of the provided information subject to the constraint B. This case can again be modeled by a combinatorial optimization problem. Specifically, let again $I(i) \in \{0, 1\}$ be the binary indicator variable for selecting provider i, $i = 1, \ldots, N$, and $\mathbf{I}$ the corresponding vector. Thus, the formulation of the optimization problem in this case will be:

Problem $\mathbf{\Pi_1}$: For $I(i) \in \{0, 1\}$, $i \in \{1, \ldots, N\}$,

$$\text{maximize } r_s\big(q_p^{\mathbf{I}}; q_d\big), \text{ such that } \sum_{i=1}^{N} I(i) \cdot c_i \leq B, \qquad (15)$$

where $r_s\big(q_p^{\mathbf{I}}; q_d\big)$ is the relevancy of a "super-provider" with a QoI function aggregated from the providers indicated by selection vector $\mathbf{I}$ (as discussed earlier in relation to (14)), and defined on $R_p^{\mathbf{I}}$ in (11).

Note that in problem Π_1, the increase of the relevancy when selecting provider i does not only depend on the relevancy that provider i contributes in isolation, but depends on the providers that have already been selected prior to provider i. This is because of the possible spatial overlap between the QoI coverage region of provider i and that of the regions of the providers already selected. In the case that the providers already selected are offering good enough quality on points ω in R_p^i, adding provider i may not increase the relevancy attained at all.

Problem Π_1 is a generalization of the 0-1 *knapsack* problem [17] where the value of each item is a function of the items already selected to be included in the knapsack, as just discussed; think of the case where adding a lighter in the knapsack may reduce (even to zero) the value of adding a box of light-matches in the knapsack later on. This is captured by the use of $q_p^{\mathbf{I}}$ as the aggregate QoI p-function parameterized on the selection vector $\mathbf{I}$.

The 0-1 knapsack problem is an *NP*-hard optimization problem which means that there is no algorithm that finds the optimal solution in polynomial time. The *greedy* algorithm would need to check all 2^N different combinations between the N providers, prune those that do not satisfy the available budget and then choose the combination that maximizes the aggregate relevancy. A *dynamic programming* algorithm has been proposed that solves the problem in *pseudo-polynomial* time. The algorithm splits the main problem into smaller subproblems and stores some of the intermediate results for later use to speed up the calculation of the main problem [17]. However, this algorithm can not be directly applied to Π_1 since the aggregated spatial relevancy when adding one provider depends also on the selection of other providers, as explained earlier. Nonetheless, Algorithm 2 has been developed to solve Π_1 using the same idea of storing intermediate results.

As a dynamic programming algorithm, Algorithm 2 is trading memory space for time. In other words, it splits the problem into smaller subproblems, stores their solutions into memory, and, then, uses them to calculate the solution of the main problem. Algorithm 2 iteratively constructs the $N \times B$ matrix $\mathbf{Values} = \big[\mathit{Values}[i, b]\big]$, where is the maximum aggregate spatial relevancy of the first i providers for a budget b; the corresponding provider selections reside in the indicator vector $\mathbf{I}_i^b$. Entry $\mathit{Values}[N, B]$ stores the maximum aggregate spatial relevancy of all providers for budget B, which is the optimal solution for Π_1 and the optimal provider selection will reside in the vector $\mathbf{I}_N^B$.

As mentioned earlier, Π_1 is an extended 0-1 knapsack problem with variable item value. Therefore, lines 7–11 of Algorithm 2 calculate the spatial relevancy (i.e., the "value") of the specific selection vector $\mathbf{I}$. The spatial relevancy of vectors $\mathbf{I}$ that have already been calculated at earlier iterations are evoked from memory. This has a significant impact in accelerating the algorithm. Moreover, lines 12–19 of the algorithm determine whether selecting a new provider will result in higher aggregate spatial relevancy (in which case the provider is indeed selected), or not.

1) Algorithm complexity: The dynamic programming algorithm for the 0-1 knapsack problem has complexity of $O(nB)$, where n is the number of items and B the available budget. In the worst case, Algorithm 2 will calculate the spatial relevancy $r_s\big(q_p^{\mathbf{I}}; q_d\big)$ at each iteration, which needs $O(N)$ time. Therefore, the absolutely worst case time complexity of Algorithm 2 is $O(N^2 B)$, where N is the total number of providers. Regarding the memory requirements, in the worst

Algorithm 2 – Budget Constrained Aggregate Relevancy

```
Initialize: $Values[0,b] = 0,\ b = 0,\dots,B$;
for $i = 1$ to $N$ do
  for $b = 0$ to $B$ do
    if $c_i \le b$ then
      $\mathbf{I} = \mathbf{I}_{i-1}^{b-c_i}$; where: $\mathbf{I}_0^{b-c_i} \overset{\text{def}}{=} \mathbf{0}$ and $\mathbf{I}_{i-1}^0 \overset{\text{def}}{=} \mathbf{0}$;
      $I(i) = 1$;
      if $r_s\left(q_p^{\mathbf{I}}; q_d\right)$ not calculated then
        Calculate $r_s\left(q_p^{\mathbf{I}}; q_d\right)$ using (5);
      else
        Get $r_s\left(q_p^{\mathbf{I}}; q_d\right)$ from memory;
      end if
      if $r_s\left(q_p^{\mathbf{I}}; q_d\right) > Values[i-1,b]$ then
        $Values[i,b] = r_s\left(q_p^{\mathbf{I}}; q_d\right)$; $\mathbf{I}_i^b = \mathbf{I}$;
      else
        $Values[i,b] = Values[i-1,b]$; $\mathbf{I}_i^b = \mathbf{I}_{i-1}^b$;
      end if
    else
      $Values[i,b] = Values[i-1,b]$; $\mathbf{I}_i^b = \mathbf{I}_{i-1}^b$;
    end if
  end for
end for
```

case, it is necessary to store the matrix **Values** of size $N \times B$, the relevancy values $r_s\left(q_p^{\mathbf{I}}; q_d\right)$ for each selection vector $\mathbf{I}$, which are in total $\min\{2^N, N \times B\}$, and the optimal selection vector $\mathbf{I}_i^b$ of size N for the $N \times B$ iterations of the algorithm.

Based on the specifics of a usage scenario, the execution of the algorithm can be accelerated both in time and memory requirements in two ways. First, recalling the discussion about the minimum bounding rectangle in the previous section, instead of examining all N providers, the algorithm can be run only for those providers whose QoI coverage region intersects with the region from the desired QoI d-function. The intersection operation will run only once at the beginning of the process and can be implemented in linear time. Then, instead of iterating for all values in the range $[0, B]$, we can calculate the *greatest common divisor gcd* of c_i, $i = 1, \cdots, N$ and B and then run the algorithm in the range $[0, B/gcd]$ with costs c_i/gcd, $i = 1, \cdots, N$.

VI. Numerical Results

The numerical results in this section were derived using a combination of MATLAB-based computations and simulations. We first compare the effectiveness of the two approximation methods of QoI functions in calculating the spatial relevancy of a single provider and, then, consider the multi-provider case and the performance of the two algorithms presented in Section IV.

A. Single-provider Spatial Relevancy

The objective of the single-provider study is assessing the robustness of the two finite-size approximation methods of QoI function descriptions in ordering providers according to their relevancy to a desired QoI function.

Due to the ease by which they can flexibly approximate several shapes with respect to orientation, flatness, peak(s), etc., we constructed QoI d- and p-functions using mixtures (superposition) of Gaussian density functions.[3] The parameters of these shapes included their relative position on the plane, their maximum value and the number of Gaussian functions mixed. These functions were approximated using the sampling and B-spline methods with M parameters (see Section IV). Then, these approximations were used to calculate the relevancy of providers based on each method using expression (8) and order the providers accordingly.

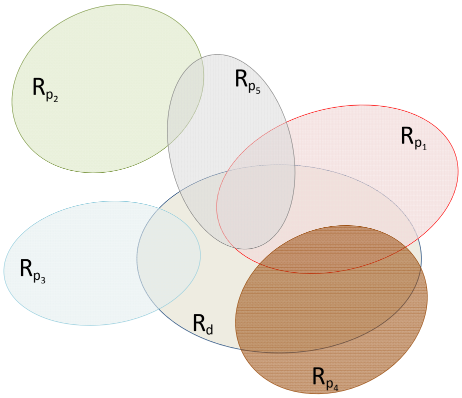
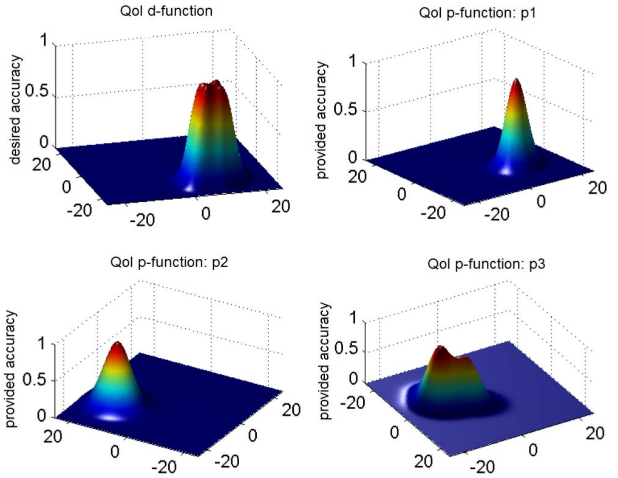


Fig. 6. The rural topology case.


QoI d-function

QoI p-function: p1

QoI p-function: p2

QoI p-function: p3

Fig. 7. QoI example d- and p-functions functions for the rural topology.

With regard to the regions R, we considered two topology cases: (a) a *rural* topology where the desired and the various provider regions are dispersed in an area, see Fig. 6; and (b) a *urban* topology where the desired and the various provider regions line-up along city streets (the "Manhattan street" topology), see Fig. 8. Figures 7 and 9 show QoI example d-

[3] The result of such mixtures does *not* posses the properties of a density function.

and p-functions corresponding to the two topology examples in figures 6 and 8. Note that Fig. 7 shows the p-functions of only 3 (p_1 through p_3) of the 5 rural providers shown in Fig. 6

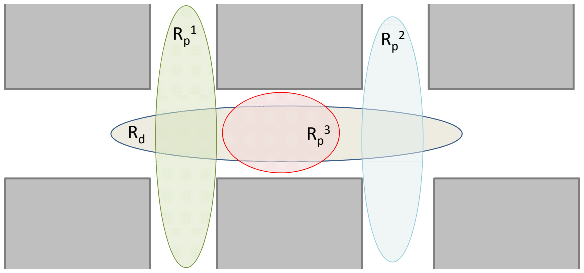


Fig. 8. The urban (Manhattan street) topology case.

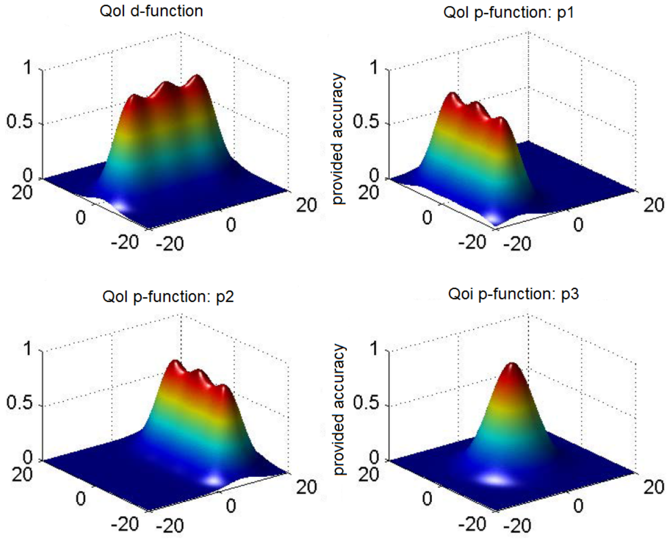


Fig. 9. QoI example d- and p-functions for the urban topology.

We applied the single-provider relevancy method to these example cases using different values of M, the maximum number of approximation parameters. Then, the spatial relevancy metric was calculated according to the two approximation methods and was compared against the actual spatial relevancy of the providers using their original QoI function. We studied: (a) the estimation error as a function of M; (b) the comparison of the two methods with respect to their effectiveness to order providers; and (c) the effect of the estimation error for each method on ordering providers according to spatial relevancy. Note that the latter provider ordering is what we are ultimately after. Specifically, the goodness of the approximation is judged not in absolute terms (which is a comparison over a continuum of values) but rather over an ordering outcome (which is a comparison over a finite set of alternatives).

The measurements presented in figures 10–12 illustrate the robustness of each method with regard to this objective. The top plots in Fig. 10 and Fig. 11 show the Spatial Relevancy of the providers calculated using the sampling method for the rural and urban topologies respectively, while the bottom plots show the respective behavior of the B-spline method. As Fig. 12 shows, the estimation error for the spatial relevancy of each provider is relatively low even when using around 100 parameters for the QoI function approximation. Comparing the performance of the two methods, Fig. 12 shows that the B-spline method yields, in general, lower approximation error despite the fact that the sampling method has smaller error for very low number of parameters in the rural topology scenario. However, the B-spline method is clearly superior in the relative ordering of providers. More specifically, there are no misordering effects even when the spatial relevancy of some providers is almost identical, as in the case of providers 1 and 2 for the urban topology case. This is indicated in the bottom plot in Fig. 11 by the fact that the red and blue lines do not intersect; intersections, such as the ones appearing in the sampling method case in the top plot, would mean a change in the relative order of provider relevancy.

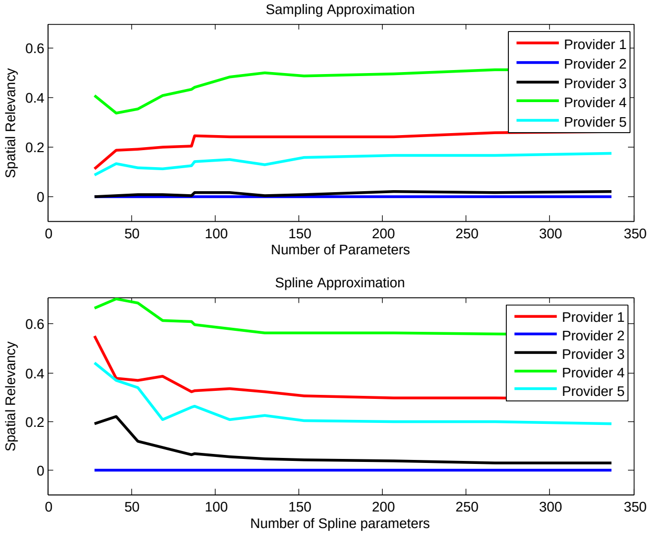


Fig. 10. Spatial relevancy for the rural topology.

As previously mentioned, we have used (8) to compute the provider relevancy. A couple of comments are in order regarding these computations. The intersection of the desired and provided regions was computed using the fast algorithm to determine the intersection of convex polygons described in [18]. Due to the requirement for convex regions, the *convex hull* of (the generally) non-convex R regions was calculated before applying the algorithm. However, this operation does not affect the value of the spatial relevancy.

The sampling approximation and the spatial relevancy metric in (8) involved uniform sampling within the regions R of the QoI functions and the reconstruction of the estimated values was done using linear interpolation. Both uniform sampling and linear interpolation processes were made using functions provided by MATLAB. Likewise, the B-splines approximation and the spatial relevancy metric in (8) involved a sampling process of the continuous QoI function. We made

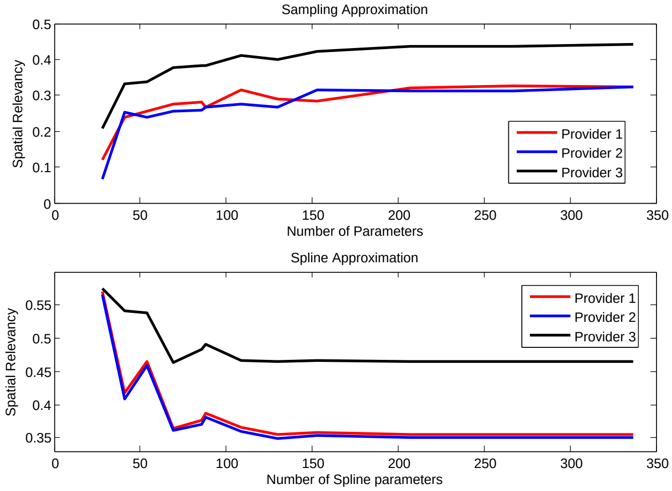


Fig. 11. Spatial relevancy for the urban topology.

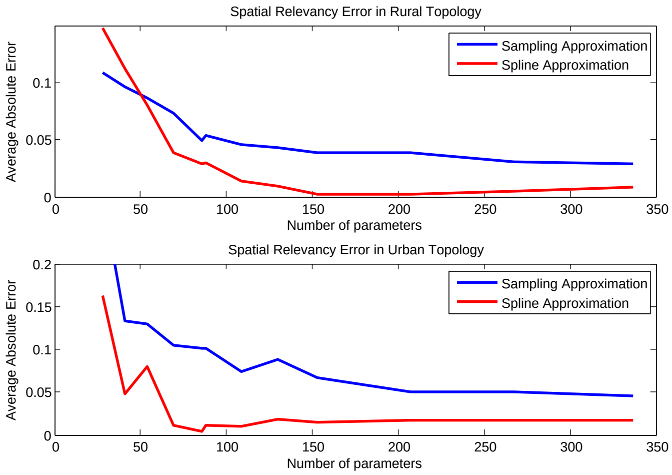


Fig. 12. Error Comparison of Approximation Methods.

use of the B-spline generation algorithm provided in MATLAB which minimized the squared-error at the sampling points; these were uniformly spaced on the respective regions R. This uniform sampling technique used during the two scenarios gave sufficiently good approximations even for the case that the special relevancy of providers was almost identical. However, as with most approximation methods, we would expect that non-uniform sampling would have improved performance especially for QoI functions that experience regions of significant and/or abrupt changes. An example of such a case would be a QoI function for which, given a region R, $q(\omega) = 1$ for $\omega \in R$ and 0 otherwise. In such a case, dedicating more samples around the boundaries of region R would yield better results.

Note that testing the performance of the spline-generation algorithms themselves is beyond the scope of this paper. We chose B-splines as a flexible, convenient and well-studied means to address our problem of describing QoI functions with finite, expansion-proof collection of parameters. Through our analysis study we confirmed that they are also a very effective aid in ranking relevant providers compared to the standard sampling method.

Building upon this procedure of calculating spatial relevancy for single providers, in the next subsection we will present the simulation results of the two algorithms proposed for the multi-provider composition problem. Since the B-spline method proved to be more robust in ordering providers based on their relevancy, the simulation results for the multi-provider case were obtained using only the B-spline method.

B. Multi-provider Spatial Relevancy

The two algorithms proposed to solve the *multi-provider composition* problems with or without the budget constraint were also simulated in a MATLAB environment. Again, the QoI functions used were mixtures of a varying number of Gaussian density functions, randomly scaled and placed on the two-dimensional plane. Fig. 13 shows an example case, where the desired QoI function is colored in blue, and 9 providers are colored in red, cyan and green.

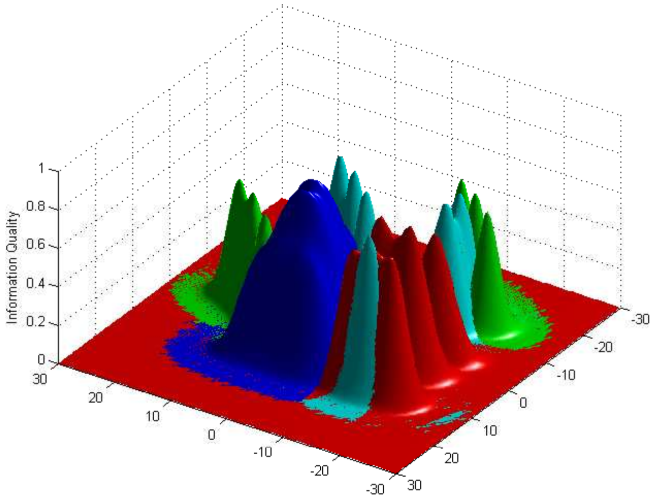


Fig. 13. QoI example functions for the multi-provider case.

The proposed algorithms are based on pseudo-polynomial heuristics to solve *NP-Hard* problems. These algorithms were adjusted to accommodate our objectives regarding the spatial relevancy of providers. Hence, the objective of our simulation study was the assessment of their effectiveness in selecting the right providers that satisfy problems Π_0 and Π_1 in Section V. The assessment is performed by comparing the solutions and execution time of the proposed algorithms against those from the exhaustive search algorithm. For the no-cost case, the latter calculates the spatial relevancy of all $(2^N - 1)$ different combinations between the N providers and the selection of the best one according to the conditions (12) of problem Π_0. For the budget constraint case, the exhaustive search algorithm includes the comparison of all *feasible* combinations, i.e., those with a total cost less than or equal to the budget, and the selection of the optimal one among them according to the conditions (15) of problem Π_1.

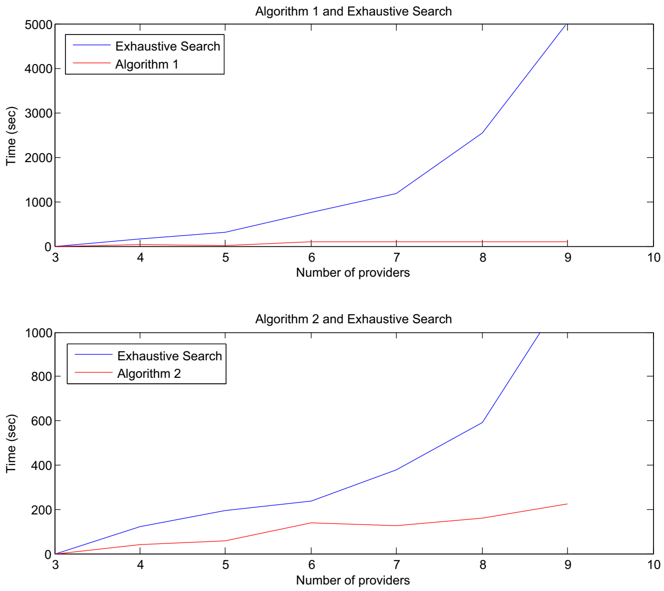


Fig. 14. Execution times comparison between algorithms 1, 2 and exhaustive search using a 2.4 GHz dual core Windows PC with 4 GB of RAM.

Fig. 14 shows the comparison of the execution time between the proposed algorithms and the exhaustive method in each case. For all cases studied, the solutions that the proposed algorithms arrived at were the same as the ones given by the exhaustive search methods, which of course are the optimal ones. As expected, the execution time of the exhaustive algorithms increases exponentially as the number of providers increases, while algorithms 1 and 2 need almost linear time. The execution of the proposed algorithms has also been accelerated by a mechanism of pruning providers not intersecting with the desired QoI function. In such cases, these providers are removed from the rest of the process with the result of further reducing the number of combinations examined.

VII. Related Work

Spatial aspects for sensor networks have been studied in many occasions in the past. Even though not sufficiently aligned with the pursuits in our work, these past studies inspired and influenced our work. For example, issues related to deployment strategies for effective spatial coverage of sensor networks are highlighted in [19], [20], [21] and references therein. Moreover, supplementing our own cited work on QoI, [22] discusses quality metadata describing geospatial information. Ref. [8] provides an extensive review of the models for spatio-temporal information databases and related queries. Ref. [23] considers spatial and thematic relevancy for matching documents to queries focusing in document ranking. Ref. [24] considers the problem of selecting the appropriate battery-operated sensors in order to maximize the life-time of the network based on the spatio-temporal relevancy correlation of the measured information between sensors in the same area. Similar problem is also investigated in [25] regarding in-network data aggregation of spatially and temporally correlated information generated by neighboring sensors.

Ref. [26] describes a process for merging topological maps where the possibility of the unbounded increase of metadata/parameters becomes evident. Granted, our case is not equivalent to merging topological maps, yet the underlying problem of metadata explosion still exists whenever we compose behaviors (the QoI functions) defined over different spatial horizons; see also [27] which deals with building and manipulating maps described by simple rectangles. Ref. [28] considers summarizing 2D shapes via a bounded number of parameters. These shapes could correspond to our regions R and, thus, the proposed approach in [28] could serve as an alternative to our B-spline approach. We do not discount the latter approach and could have been used in our paper as well. However, given that we ultimately pursue a comparison and selection of relevant providers, we found the use of the B-spline approach more flexible. Finally, our inspiration in using splines comes from [29] which considers the explosion of time-decaying security metadata of documents produced by the combination of contributing documents.

Our work examines a distinctively different aspect for sensor networks from the aforementioned studies. It is concerned with the operational aspects of information consumers dynamically selecting information providers (possibly representing multiple sensor networks) under the novel context of spatially varying interests and capabilities of the consumers and providers, respectively, while considering both coverage and QoI aspects.

VIII. Summary

In this paper, we introduced a novel problem area for sensor networks that of identifying and selecting sensory-information *relevant* providers based on their sensing capabilities in relation to an application's information needs along spatial dimensions. This problem will become more and more prominent as the number of providers increases and their sensing capabilities change spatially, such as when using wireless and mobile sensor networks operating over a multi-administrative domains, e.g., vehicle-mounted sensors, participatory sensors, etc.

Within this area we derived a relevancy metric based on the concept of QoI functions that describe the desirability or quality levels of the information desired or produced at a given location. We then developed a finite, expansion-proof technique based on two methods: sampling- and spline-based approximations to describe QoI functions and use these to advertise desired and provided sensing capabilities. The use of expansion-proof descriptors rises from the need for predictable resource usage (e.g., storage, communications) especially when considering composite providers built from the aggregation of other "regional" providers. Finally, we have formulated related optimization problems and proposed efficient algorithms for selecting the best single, or multiple collection of providers that are most relevant to our needs given various constraint objectives.

Future work in this novel area, may include the study of the various architectural aspects related to QoI function advertisements eluded earlier in the paper, temporal extensions to accommodate time-varying QoI functions that could result by system impediments, such as loss of sensors, and fluidity of sources such as in participatory sensing.

References

[1] N. Gershenfeld, R. Krikorian, and D. Cohen, "The Internet of Things," *Scientific American*, October 2004.

[2] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *World Sensor Web Workshop (in ACM Sensys'06)*, Boulder, CO, USA, Oct. 31 2006.

[3] C. Bisdikian, J. Branch, K. K. Leung, and R. I. Young, "A letter soup for the quality of information in sensor networks," in *IEEE IQ2S PerCom Workshop*, Mar. 2009.

[4] C. Bisdikian, L. M. Kaplan, M. B. Srivastava, D. J. Thornley, D. Verma, and R. I. Young, "Building principles for a quality of information specification for sensor information," in *12th Intl Conf. on Information Fusion (FUSION'09)*, Seattle, WA, USA, July 2009.

[5] G. Tychogiorgos and C. Bisdikian, "Selecting relevant sensor providers for meeting your quality information needs," in *12th IEEE Int'l Conf. on Mobile Data Management (MDM'11)*, Lulea, Sweden, June 6–9 2011.

[6] C. H. Liu, C. Bisdikian, J. W. Branch, and K. K. Leung, "Qoi-aware wireless sensor network management for dynamic multi-task operations," in *7th IEEE Conf. on Sensor Mesh and Ad Hoc Communications and Networks (SECON'10)*, Boston, MA, USA, June 2010.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006.

[8] N. Pelekis, B. Theodoulidis, I. Kopanakis, and Y. Theodoridis, "Literature review of spatio-temporal database models," *The Knowledge Engineering Review*, September 2004.

[9] J. J. Benedetto and P. S. G. Ferreira, *Modern Sampling Theory: Mathematics and Applications*. Boston, MA, USA: Birkhauser, 2001.

[10] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 2nd ed. New York, NY, USA: Macmillan, 1992.

[11] E. W. Cheney, *Introduction to Approximation Theory*. New York, NY, USA: Mc-Graw-Hill, 1966.

[12] H. Prautzsch, W. Boehm, and M. Paluszny, *Bezier and B-Spline Techniques*. Springer-Verlag, 2002.

[13] C. de Boor, *A Practical Guide to Splines*. Springer, 2001.

[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. MIT Press, 2009.

[15] R. Karp, *Reducibility Among Combinatorial Problems*. Plenum Press, 1972.

[16] J. Kleinberg and E. Tardos, *Algorithm Design*. Addison Wesley, 2005.

[17] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, 1990.

[18] J. O'Rourke, C.-B. Chien, T. Olson, and D. Naddor, "A new linear algorithm for intersecting convex polygons," *Computer Graphics and Image Processing*, pp. 384–391, 1982.

[19] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava, "Coverage problems in wireless ad-hoc sensor networks," in *in IEEE INFOCOM*, 2001.

[20] C. fu Huang and Y.-C. Tseng, "The coverage problem in a wireless sensor network," in *in WSNA*. ACM Press, 2003.

[21] A. Ghosh and S. K. Das, "Coverage and connectivity issues in wireless sensor networks: A survey," *Pervasive and Mobile Comp.*, June 2008.

[22] R. Devillers, Y. Bdard, and R. Jeansoulin, "Multidimensional management of geospatial data quality information for its dynamic use within GIS," *Photogrammetric Engineering & Remote Sensing*, February 2005.

[23] B. Yu and G. Cai, "A query-aware document ranking method for geographic information retrieval," in *4th ACM Workshop on Geographical Information Retrieval (GIR'07)*, Lisbon, Portugal, Nov. 9, 2007.

[24] P. Basu, A. Nadamani, and L. Tong, "Extremum tracking in sensor fields with spatio-temporal correlation," in *MILCOM'10*, Oct. 31–Nov. 3 2010.

[25] A. Deligiannakis and Y. Kotidis, "Geosensor networks," S. Nittel, A. Labrinidis, and A. Stefanidis, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. Exploiting Spatio-temporal Correlations for Data Processing in Sensor Networks.

[26] W. H. Huang and K. R. Beevers, "Topological map merging," *The Int'l J. of Robotics Research*, August 2005.

[27] W. Xue, Q. Luo, L. Chen, and Y. Liu, "Contour map matching for event detection in sensor networks," in *ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'06)*, Chicago, IL, USA, June 2006.

[28] J. Hershbergery, N. Shrivastava, and S. Suriz, "Summarizing spatial data streams using clusterhulls," in *8th Wksp on Algorithm Engineering and Experiments (ALENEX'06)*, Miami, FL, USA, Jan. 2006.

[29] M. Srivatsa, D. Agrawal, and S. Reidt, "A metadata calculus for secure information sharing," in *16th ACM Conference on Computer and Communications Security (CCS'09)*, Chicago, IL, USA, Nov. 2009.