Making mobile users' devices aware of the surrounding physical environment: an approach based on cognitive heuristics

M. Mordacchini and A. Passarella
Istituto di Informatica e Telematica
CNR, Italy

M.J. Chorley, G.B. Colombo and V. Tanasescu
School of Computer Science & Informatics
Cardiff University, UK

***Abstract*—In this paper we investigate the properties of a cognitive heuristics based approach, by way of which the mobile devices of users can become aware of the physical environment surrounding them. Cognitive heuristics are the mental models used in cognitive psychology to describe how human brains efficiently process the huge amount of information constantly coming from the environment around us. As personal mobile devices represent proxies in the cyber world of their human users, we investigate how the same cognitive heuristics can be used by mobile devices to become self-aware of the features of the physical environment around their users. Specifically, we assume that physical locations are described as a network of tags. We consider a self-organising opportunistic network environment, where devices exchange information when meeting directly. We propose algorithms based on cognitive heuristics through which users' nodes obtain tags either directly when coming in proximity of locations, or indirectly through other nodes they meet. We analyse the properties of the networks of tags resulting at individual nodes, as they emerge from this process, as a function of various cognitive parameters. We show that using cognitive heuristics leads, under the same resource constraints, to much more effective information diffusion with respect to other reference solutions. Interestingly, we find critical thresholds for key parameters that discriminate between information diffusion and information loss. Finally, we show that, despite resource constraints, the structure of the network of tags at individual nodes is remarkably close to the ideal that would be obtained with infinite resources.**

I. Introduction

The ubiquitous and pervasive presence in the physical world of devices that interact among themselves, their users and other sources of information in the environment is leading to an increasingly complex information landscape, where data flows from the physical world to the cyber one, and vice-versa. This scenario is known as the *Cyber-Physical World* (CPW) convergence [1].

In this scenario, the opportunistic networking paradigm [2] is considered as one of the enabling paradigms for a wide range of applications, including smart cities, e-health, intelligent transportation systems, etc [1]. In this context, it will play a key role in making the network self-organised at multiple levels. At the networking level, opportunistic networks exploit the store-carry and forward paradigm, and opportunistically exploit direct encounters between nodes to carry messages towards destinations. In a data-centric perspective, the opportunistic paradigm allows nodes to self-organise in order to disseminate data opportunistically, for example on the basis of the users interests. In addition, the role of mobile devices in the CPW convergence scenario is particularly relevant, since they are actually *proxies* of their users in the cyber world. Acting as *proxies* of their human users, they are in charge of autonomously discovering, collecting and evaluating the information available in the cyber world, determining its relevance for their users and taking the data that is of interest for them. This situation is very similar to what human brains constantly do, both in isolation, and during interactions (i.e., discussions) between people. Indeed, humans are continuously presented with a vast amount of information coming from the physical environment they are acting in. Brains are able to swiftly react and process new information, by asserting its relevance with respect to the comprehension and perception of the surrounding environment. Human brains are able to perform this task in spite of limits of time and knowledge thanks to the human ability to organise the information in memory and by using simple cognitive decision-making rules, known as *cognitive heuristics* [3]. These are very effective, yet simple and "computationally inexpensive" rules, that are functionally described in the cognitive science literature.

In this paper, we propose to apply human memory organisation models and information selection rules (i.e. cognitive heuristics) in order to let mobile devices become aware of the features of the physical environment where they move. Using cognitive heuristics at nodes is motivated exactly by the role of proxies of users personal devices, that, therefore, organise and process information in the cyber world as they human users do in the physical world. According to the proposed algorithms, users personal devices self-organise through direct interactions among them and with physical locations in the environment, in order to efficiently disseminate information about the physical locations towards relevant users. As a side effect of the dissemination process, nodes become aware of the features of the environment their users are moving in. More specifically, we assume that locations in the physical world spread their description in the environment, in the form of sets of tags.[1] Mobile devices passing near those locations can then be exposed to this information and can interact with physical locations in order to acquire their tags. Moreover, mobile nodes communicate with other mobile devices and opportunistically exchange information, thus mutually increasing their knowledge about the environment. In order to model the organisation of this information, we use a solution inspired by the *Semantic Associative Network models* of human memory, described in the cognitive psychology field. In these networks,

[1] Interestingly, as a case study we consider a case where these descriptions emerge from a flow of information from the physical world to the cyber one. In fact, we assume that they are extracted from online descriptions of the locations given by human users that physically visited them (see Sec. IV).

vertices represent the *semantic concepts* (i.e. tags) associated to physical locations, and the edges represent the relationships between tags (e.g. they are used to describe the same location). Starting from this structure of information, and assuming that limited resources are available upon contacts between nodes or between a node and a location (such that only limited amount of information can be transferred), a physical location or a mobile node selects, upon contact, the most relevant data to communicate to the other interacting party. In our solution, the selection of the information to be exchanged is also driven by cognitive models of how humans exchange information during a discussion. Intuitively, common concepts of interest drive the selection of the concepts that more easily come to mind, since they are more related with the common concepts. To replicate this process, we propose to exploit the *fluency heuristic* [4] strategy. The fluency heuristic is the cognitive strategy that allows the brain to choose among two or more alternatives. Among all the alternatives, this heuristic favours the ones that are recognised, i.e. have been "seen" in the environment a sufficient number of times. Among them, it then chooses the one that is perceived as being recognised faster, , because more strongly related to common concepts of the discussion. We show how the graph information representation based on associative network models and the information selection strategy based on the fluency heuristic can be coupled to achieve an effective dissemination of the physical world description among mobile nodes in an opportunistic network. We report results on the analysis of the emerging properties of graphs with tags of physical locations stored at each node as a side effect of these interactions, and the behaviour of the proposed solution under different settings of its main parameters. Specifically, we show that the cognitive-based dissemination scheme is more efficient than standard solutions not using cognitive heuristics. Through a sensitiveness analysis we highlight phase transitions for the dissemination of location information. We also highlight that the tag networks at individual nodes are structurally very similar to the asymptotic one, containing the union of all physical locations' descriptions, that would be collected with infinite available resources.

The rest of the paper is organised as follows: in Sec. II we report previous work about information dissemination in opportunistic networks and the extraction of representations of physical places from online descriptions. In Sec. III we present the main algorithms about the information organisation and exchange processes. In Sec. IV we illustrate how the description of the physical locations has been derived. Sec. V shows results about the properties and behaviour of the proposed system in a simulated environment. Finally, Sec. VI concludes the paper.

II. Related Work

The data dissemination problem in opportunistic networks is a sensitive issue that has been faced by many works in the literature (see [2] for a complete survey). All of them are based on "traditional" computer-science heuristic solutions. Only some recent works [5], [6], [7] start to consider solutions coming from the cognitive science field to devise simple and low resource-demanding schemes for effectively disseminating the information among nodes in an opportunistic network scenario. In scenarios were mobile devices are proxies of the human users, using heuristics that exploit the same mechanisms used by the human brain to take decisions prove to be particularly efficient. In fact, these solutions proved to be as efficient as other "traditional" approaches in disseminating the data toward interested nodes, while, at the same time, requiring much lower resources to reach this result. Independently from the adopted dissemination strategy, anyway, all those approaches do not take into consideration the semantic side of data, or, more precisely, how different pieces of information are connected to each other for humans because of their semantic relationship. To cope with this issue, in [8] an algorithm for spreading semantic information has been proposed. On the other hand, the work proposed in [9] represents, at the best of our knowledge, the first attempt for equipping nodes with cognitive-based solutions for the opportunistic exchange and diffusion of semantic information and its associated content. All these works, anyway, do not take into account any possible interaction between mobile devices and the physical environment the nodes are moving in. Moreover, also the data available in the network is assumed to be not related in any way to the physical context. Thus, by using those schemes, it is not possible for nodes in an opportunistic network to become aware of their surrounding environment. In this paper we propose to overcome these limitations, using a solution that allows nodes to acquire a knowledge about the physical space their users move. This is obtained through a semantic representation of the information about physical locations, and its spreading through a direct communication of the devices with the physical locations they encounter in the environment, and between them. To treach this goal, we show how cognitive memory representations and information selection schemes taken from cognitive psychology can be exploited to this end.

In order to communicate their features to mobile nodes passing near-by, physical locations need to generate their own descriptions. Virtual representations of place have been the topic of active research, with an understanding that, for the human mind, the geospatial environment is organised as places rather than sets of geospatial coordinates [10]. Places appear as complex thematic entities in relation with the physical configuration of the environment as well as with human cognition such as memory [11]. Places, as part of the environment, present physical or social opportunities for action or information [12], as well as for social interaction with other agents, or the absence of them [13]. Therefore, they are spatial regions that support information of significance for the agent in an environment and act as cognitive anchors, through salient features that make them useful or interesting, memorable because of past experiences, or desirable as expected loci of anticipated ones [14], [15], [16].

Online representation of places has been the object of several publications in the literature. For example, [17] suggests that if it is possible to 'visit' virtual representations of places and collect from them the information we are seeking. Tags could be the basis of a novel process for the extraction and building of digital location, place, events and semantic descriptions [18], [19] which could lead to a re-design of the concept of an urban area [20] closer to demands from parallel research areas, such as cognitive psychology and the formulation of a new notion of 'place' conceived of as encompassing meanings, sense of attachment, and satisfaction provided by the people interacting with it [21], [11]. The authors have presented a proof of concept for a keyword extraction methodology using online tips and reviews from a number of online sources, see [22] for a complete description. This has also been adopted in this current research to build a dataset to represent a number of sample venues of different characteristics and types.

Digital representation of information and opportunities for action available at places are made available notably through the use of crowdsourced information, such as reviews [23].

Mobile devices and location-based services now allow the monitoring of geographical positions in real time, thus moving from an initial adaptation of online maps and navigators towards services more oriented to provide reviews and personalised recommendations such as *Yelp* and *Qype*[2]. Other services combine location and user mobility information with a social networking component. Among those *Foursquare*, *Flickr*[3], and *Google+ Local*[4] have converged to a place representation that focuses on the individual needs of users, in terms of users being at a particular location at a particular time often making use of tags, annotations and other user generated content [19], thus differentiating from an initial representation of places as a stand-alone virtual location that appeared distant from the real user's needs [11], [17].

The importance of the use of shared keywords and 'tags' as a form of metadata in content organisation (collaborative tagging) has been widely recognised as a basis for modern suggestion and recommendation systems [24]. Free tags, as opposed to controlled vocabularies are generally preferred by users as more personal and less cognitively demanding [25], [26]. Location services can be used to mine crowdsourced data in a place and reusing keywords extracted from these descriptions (tags) to inform the user, use of this technique of summarisation is not yet widespread.

While most of this research assumes fixed data sharing platforms such as Online Social Networks, this papers focuses on the issue of disseminating physical locations information (i.e. tags generated according to the methodology above, better described in Sec. IV) in mobile social networks scenarios, where devices of the users exchange information between each other, without any mediation of fixed infrastructures.

III. Organisation and dissemination of physical location information

In this section, we show how human cognitive information collection and selection schemes can be exploited to drive the dissemination of the information describing locations of the physical world among mobile devices, making them (and, in turn, their users) aware of the features of the physical environment they are moving in.

We face two main challenges in designing a system for disseminating location tags. The first one is how tags are organised in the nodes' memories and associated to locations. In particular, we wish to let each device become aware that tags describing the same location are in some way related to each other and that different locations can share common sets of tags, allowing users to apply further strategies, e.g. the computation of similarities between locations, based on the locations' descriptions. The second challenge is how tags are efficiently passed either from locations to users' devices, or between users' devices, upon encounters. In the following we separately address these two challenges.

A. Memory Organisation

To address the first challenge, we refer to the cognitive associative memory description of the human brain. The human associative memory is modelled in the cognitive sciences using the so-called *Semantic Associative Network* (SN) models [27]. These models focus on the patterns and strength of associative

[2]http://www.yelp.co.uk/, http://www.qype.co.uk/

[3]https://foursquare.com/; http://www.flickr.com/

[4]http://www.google.com/+/learnmore/local/

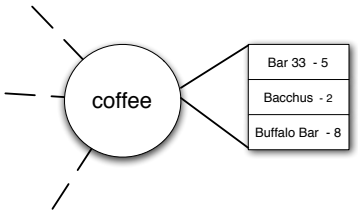


Fig. 1: Example of an annoted vertex in a user SN

linkages among concepts in the brain. Associative network models represent the memory as a graph, where concepts are the nodes. Each pair of concepts is connected in case the brain has made an association between them. Edges of an SN are weighted, where the weight reflects the strength of each association in memory.

In order to take advantage of this cognitive mechanism, we define the physical locations description and each user's memory about them as a *weighted graph* $G = \{V, E, m(e,t)\} : t \in T$. In this definition, t is an instant in the time frame T, V is the set of vertices (i.e. locations' tags) and E is the set of edges (i.e. the connections between tags). Moreover, we consider that each node (tag) of the graph G is annotated with the set of physical places associated to that tag known by the device, as illustrated in Fig. 1. Clearly, physical locations maintain only the associations with their own name. On the other hand, along with the association with the names of physical locations, mobile devices keep track of the number of times this association has been observed in exchanges with other peers. This information will be used in the data exchange process, as explained in the second part of this section.

As we stated above, in a SN, links have an associated weight that represents the strength with which the brain is able to "recall" the association between the pair of concepts in memory. This process is replicated in G using a *memory weight function* $m(e,t)$. We define this function in the following way. We consider that an edge e_{ij} has the initial condition $m(e_{ij}, t_0) = 1$, where t_0 is the edge's creation time. We consider that physical locations do not change their description over time. Thus, in the SN of a physical place, $m(e_{ij}, t_0) = 1$, $\forall e \in E$ and $\forall t \in T$. For a mobile node, at any instant in time $t > t_0$, $m(e_{ij}, t)$ decreases exponentially depending on the length of the interval (t', t) , where t' is the last time e_{ij} was "refreshed" in memory (i.e. used in interactions with other nodes or locations). We can then define $m(e_{ij}, t)$ as:

$$m(e_{ij}, t) = e^{-\beta_{ij}(t-t')} \tag{1}$$

where β_{ij} is a factor that regulates the "speed of forgetting", defined as

$$\beta_{ij} = \frac{\beta}{p_{ij}^t}$$

where β is a speed regulator parameter and p_{ij}^t is the "popularity" of edge e_{ij}, i.e. the number of times it was used during the encounters of that specific user with other people until time t. This information is stored as an additional label (together with the weight) associated with the edge, and updated as described in the following section. The exponential forgetting function is a well-known representation of the forgetting process in cognitive psychology [4]. Rather than a limit, the forget process helps human brains to discard less relevant information when making decisions [4]. For this reason, whenever the value of the *memory weight function* for an edge e_{ij} goes below a *memory weight threshold* M_{min}, e_{ij} is removed from G. Since human memories are more likely to drop information

that is rarely accessed than frequently used data, we bind the definition of the *memory weight function* to the edge popularity. Therefore, connections between tags that are rarely used during exchanges with other devices or locations are more easily "forgotten". The forget process can affect also the tags retention in memory. In fact, as shown in Fig. 2, a node in G is dropped from the graph in the case where it becomes isolated, i.e. it is not connected anymore to any other node in G, due to the deletion of one or more of its outgoing edges. Note that in the tag network stored at a physical location, edges still represent an association between tags, although clearly not bound to any human reasoning process, but representing how the tags representing that location are linked to each other. In Sec. V we explore the performance of the dissemination system starting from different ways of organising the individual locations' tags networks.

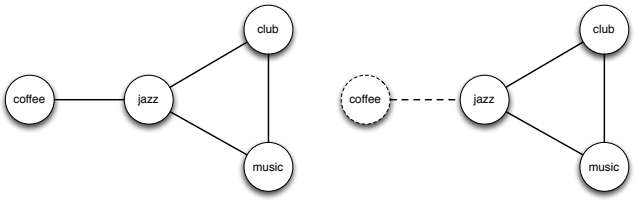


Fig. 2: Effect of edge "forgetting" on an isolated node

All the operations described above allow locations and nodes to store and handle the information in their memories. We now give a description of the mechanisms that allow the exchange and update of data using a semantic associative network description.

B. Information Selection and Exchange

We consider that, upon meeting, a node (either mobile or physical) starts to exchange its knowledge with the other party by selecting the most relevant concepts for that given contact at that time from its SN. The information exchange process starts from concepts that the two parties have in common. From those starting points, the information is selected by navigating each SN, according to a weighting scheme. This process is similar to what is called a sequential search over human associative memories [28]. This search mechanism starts with an activated semantic concept (*key-concept*) and then proceeds vertex by vertex in the SN, following the links that connect them. Whenever a "dead end" is found, the search is reinitiated. In order to determine the relevant paths to follow when exploring a SN, we apply a very simple rule: when a node in a SN has more than one outgoing edge, the link with the "strongest activation" (measured by the retrieval weight index explained in the following) is selected. This behaviour is the implementation of the Fluency Heuristic (FH) [4] cognitive strategy. We now describe the precise algorithm by which the FH is used by a physical location or a mobile node to retrieve the most relevant semantic information to be exchanged from its memory. In the following, the description is made from the viewpoint of a location or a node that is selecting the information to pass. We call this node the *donor* node, while the other party is called the *recipient* node. As we already stated, we assume that physical locations do not change their descriptions. Thus, they are always only *donor* nodes. On the other hand, two mobile nodes swap roles (donor and recipient) upon contact to realise a bidirectional exchange of information. In this phase, we call donor node's SN $G = (V, E, m(e,t))$ the *donor network*, while the other peer's semantic network $G' = (V', E', m'(e',t))$ is termed the *recipient network*. The subgraph $C = (\bar{V}, \bar{E}, \bar{m}(\bar{e},t))$ selected from the donor network to be passed to the recipient one is called the *contributed* network. The following description assumes also that there exist resource consumption constraints that limit the number of exchangeable tags (i.e. nodes of the SN) to a maximum value T_{max}. For the same reason, we assume that the donor node can select no more than L_{max} names of physical locations associated with each tag it includes in the contributed network. The pseudo-code of the subsequent description is given in Alg. 1 and 2. The pseudo-code of Alg. 1 is the same we use in [9] and we report it here for the reader's convenience. As we stated above, we assume that the dialogue starts from a set of *key vertices* $K = \{v_k | v_k \in V \cap V'\}$ (line 3 of Alg. 1), i.e. a set of semantic concepts that the SNs of the donor and recipient nodes have in common. Given this set of common concepts, the visit of the SN starts from the most relevant nodes in the set. Vertex relevance is computed by summing up the memory weights of a vertex's incoming edges (line 8). In order to represent the relevance of a vertex, we consider that it is increased every time it is included in K during information exchanges by augmenting the popularities of all the edges attached to it (lines 5–7).

Algorithm 1 Contributed Network computation at time t^*

1: Let $G = (V, E, m(e,t))$ be the donor network;
2: Let $C = (\bar{V}, \bar{E}, \bar{m}(\bar{e},t))$ be the contributed network;
3: Let K be the set of *key vertices*, $K \subseteq V$
4: **for** each $v_i \in K$ **do**
5: **for** each $e_{ij} \in E$ **do**
6: increase popularity of e_{ij}
7: **end for**
8: Let $rel_{ij} = \sum_{e_{ij} \in E} m(e_{ij}, t^*)$
9: **end for**
10: **for** each $v_i \in K$ taken in desc. order w.r.t. rel_{ij} **do**
11: $C \cup = visit(v_i, 1, t^* - t)$
12: **end for**
13: Send C to the other node

Taking the key vertices (sorted by relevance) one at a time, edges and vertices are visited and passed from the donor network to the contributed one using Alg. 2, based on the FH. Before exploring the SN, since vertices are annotated with the physical location names, once a vertex is selected, its set of annotations is considered (line 5 of Alg. 2). While physical locations always only associate their names to passed vertices, we assume that mobile nodes order vertex annotations in descending order with respect to their popularities. Thus, the locations that are most frequently associated with a given tag are chosen first to be exchanged. As already pointed out, no more than L_{max} locations are included in the annotations of the nodes included in the contributed network (lines 8–12). Then, the edges used to visit a SN are selected using the FH. Since the FH favours *recognised* objects (i.e. objects seen more than a given amount of times) against unrecognised items, we start by excluding all the edges whose popularity is below a recognition threshold θ_{rec} (line 15 of Alg. 2).

In order to replicate the subsequent discrimination made by the FH, and based on the perceived retrieval time, we consider that the most relevant edges are the ones with higher *memory weight* values and closer to a key vertex. Anyway, the longer the contact time between two nodes, the more time is available to navigate the donor network and include edges and vertices in the contributed one. These factors are taken into account in the algorithm by computing a *retrieval weight* value for each outgoing edge edge e_{ij} of a vertex v_i (line 16). The *retrieval*

weight is computed as in Equation 2:

$$w(e_{ij}, n, t^* - t) = m(e_{ij}, t^*)\frac{1 - e^{-\tau(t^*-t)}}{n} \quad (2)$$

where $m(e_{ij}, t^*)$ is the memory weight value of e_{ij} at time t^*, n is the number of hops in the shortest path to the nearest key vertex and τ is a "speed" factor that regulates the dependency of this value on the communication duration $(t^* - t)$. With a longer communication time, edges have more chances to be "warmed up" by the "discussion" and, then exchanged. We also refer to the *retrieval weight* as the *warm* value. Using this quantity, edges are sorted according to their retrieval weight and are then taken one at the time in descending order (line 16). Clearly, since for each physical location $m(e_{ij}, t^*) = 1$ for each edge e_{ij}, the retrieval weight for an edge in physical location SN depends only on the distance from a key vertex and the communication duration. Each selected edge is included in the contributed network and allows the donor network exploration from this link to continue (lines 4–21). We consider that the edges whose retrieval value is below a threshold W_{min} are not relevant for the actual information exchange and, thus, they are excluded from the contributed network (line 17). Moreover, once an edge is included in the contributed network, its memory weight is set to the maximum in both the donor and the contributed networks, since its inclusion in the exchanged data corresponds to an "activation" in memory (lines 19 and 20).

Algorithm 2 Functionc $visit(v_i, n, t^* - t)$

1: Let $G = (V, E, m(e, t))$ be the donor network;
2: Let $C = (\bar{V}, \bar{E}, \bar{m}(\bar{e}, t))$ be the contributed network;
3: **if** $|\bar{V}| <$ *tag_limit* **then**
4: Let $\bar{v}_i = v_i$
5: Let L be the set of the annotations of v_i
6: order L in desc. order w.r.t. the annotation pop.
7: Let $\bar{L} = \emptyset$
8: **if** $|L| \leq L_{max}$ **then**
9: $\bar{L} = L$
10: **else**
11: put the first L_{max} elements of L in $\bar{L}$
12: **end if**
13: Annotate $\bar{v}_i$ with $\bar{L}$
14: $\bar{V} \cup = \bar{v}_i$
15: Let $R = \{e_{ij} \in E | p_{ij}^{t^*} \geq \theta_{rec}\}$
16: **for** each $e_{ij} \in R$ in desc. order w.r.t. $w(e_{ij}, n, t^* - t)$ **do**
17: **if** $w(e_{ij}, n, t^* - t) \geq W_{min}$ **then**
18: $\bar{E} \cup = e_{ij}$
19: $m(e_{ij}, t^*) = 1$
20: $\bar{m}(\bar{e}_{ij}, t^*) = 1$
21: $C \cup = visit(v_j, n + 1, t^* - t)$
22: **end if**
23: **end for**
24: **end if**
25: Return C

Whenever $|\bar{V}| = T_{max}$ and/or no other edges can be selected from the donor network, the contributed network computation ends and the resulting graph is passed to the recipient node. This node, in turn, merges the received contributed network to the recipient one by simply adding all the missing vertices and edges. Moreover, as for the donor node, all the edges received from the contributed network (new or already present) set their *memory weight* to 1, since they are "activated" by the "conversation". When merging the contributed and the recipient networks, the recipient node also considers the annotations of each vertex of the contributed network. In case the vertex was already present in the recipient network, the recipient node increases the popularity of all the annotations that already exist in the recipient network vertex and are present in the contributed one. Otherwise (i.e. the vertex was not previously in the recipient network, or the annotation is not already present) the popularity of the received annotations is set to 1.

IV. Dataset Creation

We have produced a data set for venue representation using the methodology presented as a proof of concept in [22] and shown as example in Figure 3. This consists of a keyword extraction process from aggregated text obtained from online reviews thought as representative of the perception of users about a place (venue), rather than its objective description. The procedure returns a weighted list of keywords (tags) where each weight represents the 'importance' of specific keywords (for example its frequency in the text in the simplest version). This methodology can be described by the processes described in the following subsections. :

Fig. 3: Tag-list generation process

A. Document Aggregation

For each venue we produce a document that aggregates text from online reviews, user tips and comments, and other keywords that can be found online. Google+, Yelp, Qype (for reviews) and Foursquare (for tips) are used as data sources. Using the text document for each venue as an input the procedure that extracts (Sec. IV-B) and weight (Sec. IV-C) a list of keywords is then executed. The end result is that each venue is represented by a n-dimensional vector v with each dimension $v[i]$ being mapped to a distinct individual keyword (tag).

B. Keyword Comparison and Filtering

This is obtained by:

- Using the Natural Language Toolkit library [6] to filter by different parts of speech (POS) such as adjectives, nouns, verb and adverbs. This is also used to tokenise, un-capitalise, strip of punctuation, remove unwanted words such as conjunctions, stop-words, repeated words, non-english words etc. We have here focused on adjectives and nouns as POS's.
- Using different options in parsing the test document, for example we can include everything (e.g tips, reviews, comments, categories etc.) or include only the proper reviews (and tips or comments) but leave out the text classified as representing a venue type or category. This is the approach used to produce our data set, which only considers reviews and tips.

[6] http://nltk.org/

C. Relevance Weighting

There are different procedures to produce the tag weighting. In particular we can use:

- The word count of each term in the text document.
- More sophisticated procedures such as TF-IDF (see [29] for technical details). In this case the weighting for a specific venue also depends on documents about other venues (the collections of documents used is called the *corpus*).

Term Frequency (*TF*) is a simple weighting scheme for keywords in a document, that uses the bag of words model. *TF* assumes the weight of a keyword to be equal to the number of occurrences of term t in the document. Using term frequency alone however has little discriminating power in a themed corpus, as some keywords will probably be found in all documents. For example the word 'beer' will be found in many or all pub reviews. The idea is therefore to adjust term frequency using the count of occurrences of the term in the whole collection. This measure is the document frequency DF, or number of documents that contain a term t. To use DF to scale the term frequency, the inverse document frequency IDF of a term t is defined as

$$idf(t) = \log(N/df(t)) \tag{3}$$

where N is the number of documents in the corpus. The *TF-IDF* weighting scheme assigns to term t a weight in document d given by:

$$tf - idf(t,d) = tf(t,d) \times idf(t) \tag{4}$$

The weighting has the following characteristics:

- highest when the term occurs many times within a small number of documents
- lower when the terms occurs fewer times in a document, or occurs in many documents
- lowest when the term occurs in virtually all documents

In this work we have considered about 10 venues in the city of Cardiff (UK) using a 'global' corpus consisting of the collection of all documents considered. This results in a total of 2210 tags, with an average of 221 tags per location.

V. Properties of the semantic network of locations at mobile nodes

In this section we present results about the properties of the information dissemination process and of the SN of mobile nodes obtained in a simulated environment. Tab. I shows the main parameters used for the simulation. We considered a $1000m^2$ wide area where 100 nodes move according to a random waypoint model. Inside this area, there are 10 static physical locations (placed uniformly at random in the simulation area) that spread their information. The description of these locations is derived from a real-world dataset, described in more detail in Sec. IV. This simulation settings have been chosen as they are able to highlight the *general* behaviour and the macroscopic features of the proposed approach, allowing us to give an intial evaluation on the system performance in a realistic scenario. In order to model the SNs of the physical locations, in the following we consider three different configurations, based on the tag popularities derived from the TF-IDF frequencies of the locations tags in the dataset. The first configuration organises the nodes of the locations' SNs as a chain, with the most relevant tag at one end, connected to the second most popular tag, and so on, till the least popular tag at the other end. Semantic networks derived from online description could be also viewed as the aggregation of the associations made by a plurality of different users. Studies in the cognitive sciences (e.g. [30]) report that aggregate semantic associative networks show scale-free properties. Therefore, we also use two more clustered approaches, obtained using the algorithm reported in [31]. Since this is a growing model of a graph with scale-free properties, for each location we run the algorithm by introducing the vertices in the graph growing process on the base of their TF-IDF order of relevance. The first configuration has a clustering coefficient of about 0.2 and the other one has a clustering coefficient of 0.5. Hereafter, we refer to all these three configurations as the Chain, CC=0.2 and CC=0.5. configurations, respectively. At the start of the simulation each node SN is initialised by choosing a group of tags from the set of all the available tags. Each tag has a 0.01 probability to be added to a node SN. Initially, tags in a node SN are not connected to each other, i.e. the node SN do not have any edge. Moreover, tags are not annotated. Nodes should then acquire the knowledge about relantionships between tags and association to physical location through the interaction between physical locations and other nodes. In the following, the performance metrics that we use are the Hit Ratio and the Coverage. The first one is the average over all the nodes of the ratio between the number of tags held by each node and the overall number of tags available from the physical locations. The Hit Ratio indicates the amount of information acquired by the nodes in the system. The second measure is defined as the average of the per-node Coverage. This value, in turn, is computed as average (computed over all locations stored in the node SN as tags annotations) ratio between the the number of tags in the node SN annotated with a location, and the number of tags that describe that location and are also stored at the node. Note that the node may not store all the locations associated to a stored tag, and thus coverage measures how complete is the information for the stored tags. In order to make the values of the *memory* and *retrieval* weights more intuitive for the reader, in the following we use these conventions. A notation like $forget = 50$s means that the M_{min} weight is set in such a way that edges with popularity 1 are dropped from a SN in case they are not seen before 50s from the last time they were used in an exchange. On the other hand, a notation like $warm = 25$s means that the W_{min} threshold is computed taking into account, as a reference case, an interaction between nodes of 2s, that let to include (*warm up*) at least edges at distance 1 from a key node if they are not used (i.e. they were subject to the forget process) from no more than 25s. All the reported results are the mean of 10 different runs of the algorithm, obtained by using 10 different mobility traces for the nodes and 10 different placements in the area for locations.

TABLE I: Main simulation parameters

Simulation Parameters	
Simul. Area	$1000m^2$
Numb. of Nodes	100
Numb. of Phys. Loc.	10
Node speed	unif. in $[1, 1.86]m/s$
Transm. range	20m
Simulation time	75000s
β	0.1
τ	0.1
θ_{rec}	5

A. Comparison with an epidemic dissemination model

In this first set of results, we show a comparison between our solution and an epidemic-like [32] data dissemination scheme. In this epidemic scheme, locations and mobile devices select the tags to pass to another encountered peer uniformly at random from the set of data they hold in their memories. Upon contact, for each exchanged tag, a set of known associated locations is also passed to the other party, by selecting them uniformly at random from the ones in memory. The epidemic scheme is also subject to the same restrictions of the cognitive-based approach. Thus, no more than T_{max} tags can be passed at each encounter and no more than L_{max} locations can be associated with each exchanged tag. Moreover, we assume that data in the epidemic approach is subject to an aging process, similar to that of the cognitive case. For each tag stored by a device in the epidemic scenario, we compute a "popularity" value in the same way as for edges in the cognitive approach. Then, we can define a *memory weight* function $m(d_i, t) = e^{-\beta_i(t-t')}$ applied to each data d_i. It is simply the cognitive *memory weight* function (Form. 1) applied to data items rather then edges. We use this function with exactly the same parameters as the cognitive function, i.e. the β value and the same M_{min} threshold, used to drop items from memory. This epidemic scheme is a very simple benchmark for our algorithm. In particular, it does not use any semantic representation of tags and any association between them. Epidemic is the most simple scheme that can be used to disseminate location information without using our cognitive based approach, but subject to the same resource constraints. Comparing our scheme with epidemic allows us to check that the former, under the same resource constraints, is able to achieve better information dissemination, thus resulting in a more efficient use of the available resources.

In the following simulations, we vary the maximum number of exchangeable tags T_{max} and the memory weight threshold M_{min} values, keeping fixed the minimum retrieval W_{min} and maximum number of exchangeable vertex annotations L_{max} parameters. In particular, $L_{max} = 2$ and W_{min} is computed for a *warm time* $= 25$s. In all the results reported in Figs. 4–5, it is possibile to see that the cognitive-based approach is able to outperform the epidemic approach in terms of both the Hit Ratio and Coverage, independently of the tag and forget parameters and the cognitive location tag network configuration. Moreover, note that, in all cases the epidemic approach reaches a point where it enters a sort of oscillatory behaviour, where the values of the performance figures increase and decrease, floating around a stabilisation point. This behaviour is due to the effect of the forgetting process. In this region, the increase of the performance figures due to newly acquired tags is soon compensated by the drop of the oldest, least popular tags.

On the other hand, the cognitive-based approach takes advantage of a structured information representation, where the forget process is applied to the edges in each SN, before it impacts on the data. Specifically, organising tags in semantic networks allow the system to avoid oscillations, as important tags have time to become "well established" in the nodes semantic networks, while tags that are really less important are correctly dropped. Finally, it is possible to observe that more clustered location SNs allow a better flow of information. This is expected, as clustered tags are closer to each other in the locations SN, and thus it is easier for each tag to be passed in the mobile nodes SN (remember that the lower the number of hops between a tag and the set of common tags, the higher the chance of ending up in the contributed network). Anyway, differences between the two clustered configurations are evident only in the experiment with the lowest T_{max} and forget values (Fig. 4). It has to be noted that the Chain configuration achieves Coverage results very close to the clustered configurations when using higher forget and tag limit values (e.g. Fig. 5). Thus, even with the most difficult starting configuration, information spreads efficiently using the cognitive based schemes.

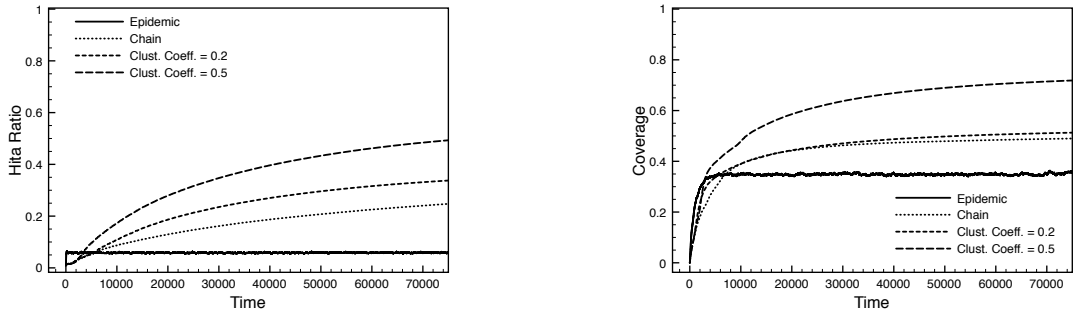
B. Sensitiveness Analysis

The previous results highlight the dependance of the performance figures on the T_{max} and forget parameters. In order to better understand this point, in the next set of experiments we report the performance of the system under various values of all its main parameters. The results are obtained for a Chain location SN configuration, that is, the most difficult case for disseminating the information. Similar outcomes, with higher performances, as observed in the previous results, can be observed with all the clustered location SN configurations. If not otherwise stated, the results are obtained with *forget* $= 75$s, $T_{max} = 150$, *warm* $= 25$s, and $L_{max} = 2$. Fig. 6a shows the Hit Ratio variations due to various values of the forget parameter. As observed in the comparison with the epidemic-like model, the higher the forget time, the higher the Hit Ratio. Note that the difference between the two lowest forget values are far bigger than the difference between the other settings. In particular, the lowest forget value leads to a very slow increase of the Hit Ratio over time. A similar situation is observed in Fig. 6b for the T_{max} value, where the lowest value results in an even slower Hit Ratio increase than for the forget case. On the other hand, fewer differences emerge for the warm value (Fig. 6c). In particular, an increase from $warm = 25$s to $warm = 50$s does not give a great difference in the final Hit Ratio values. Therefore, the system seems to be less sensitive to this value than the previous ones. Note that, although higher values of forget and tag limits could result in higher Hit Ratios, their setting should be pondered. In fact, higher T_{max} values lead to a higher resource consumption, since more information could be exchanged at every encounter, while higher forget values make the system less responsive and adaptive to changes in the environment. In fact, in case the oldest information suddenly becomes less relevant, making it stay for longer periods in memory will force the devices to evaluate it during encounters, even if it is now not relevant.

The Coverage value has a dependance to the various parameters which is similar to the dependence of the Hit Ratio. In Fig. 7a we show the impact of the forget value. As for the Hit Ratio, it is possible to see that a small increase in the forget threshold (from 50s to 75s) leads to a great increase in the Coverage measure. These observations seems to point to the fact that there exist a point over which the information dissemination process is able to proceed more rapidly. We investigate this fact more in detail in the next set of experiments. Fig. 7b shows that the higher the number L_{max} of suggested location with each tag, the higher the Coverage. Since increasing the L_{max} value produces a higher resource consumption during exchanges, the tuning of this parameter should take into account a trade-off between higher performances and resource consumption constraints, as for the T_{max} parameter.

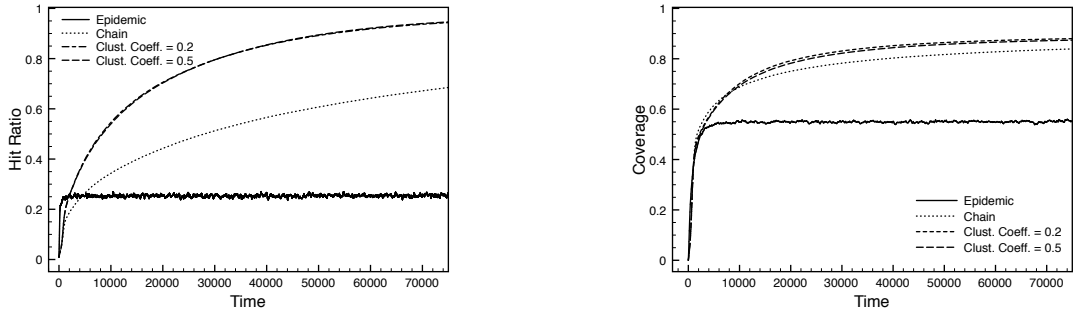
C. Macroscopic properties of the mobile nodes SNs

The following results focus on more general, macroscopic properties of the system. In particular, we investigate more



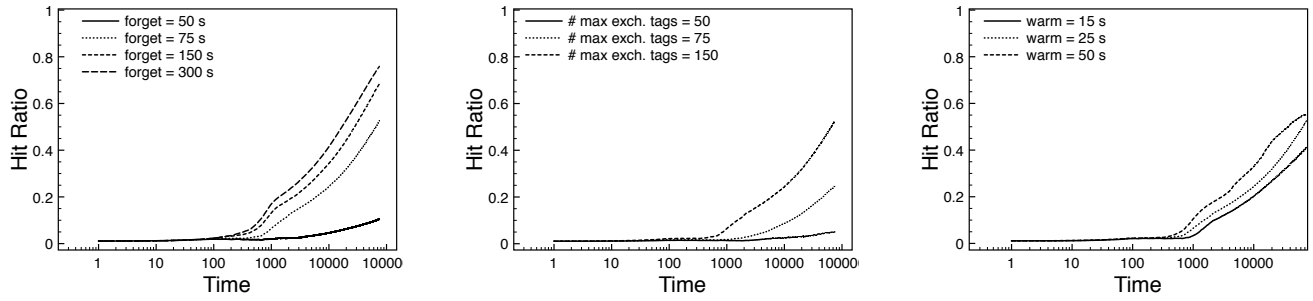
(a) Hit Ratio (b) Coverage

Fig. 4: Hit Ratio (a) and Coverage (b) results comparison; #tags =75, forget=75s



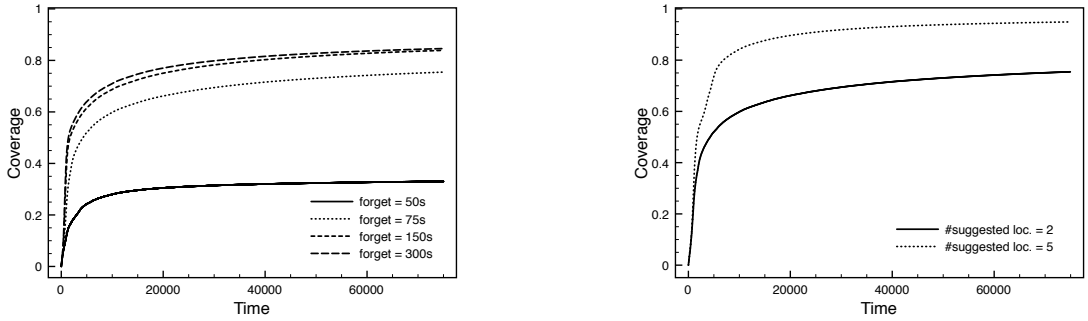
(a) Hit Ratio (b) Coverage

Fig. 5: Hit Ratio (a) and Coverage (b) results comparisons; #tags =150, forget=150s



(a) Forget impact (b) Tag Limit impact (c) Activation threshold impact

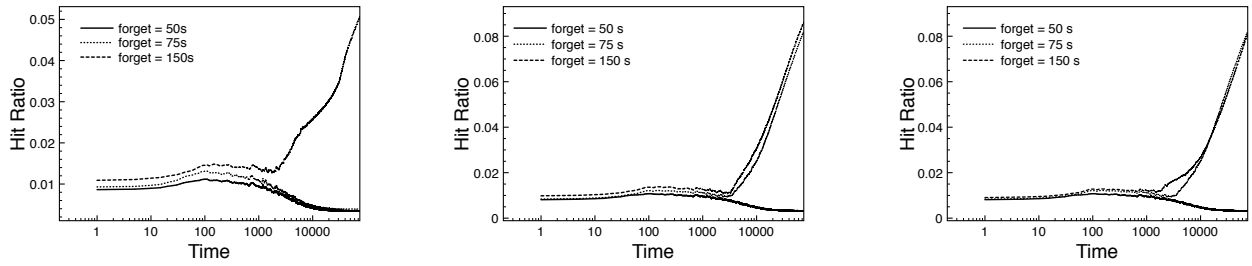
Fig. 6: Impact on the Hit Ratio of various settings of the parameters of the forget (a), tag limit (b) and warm activation (c) values



(a) (b)

Fig. 7: Impact on the Coverage of various settings of the parameters of the forget threshold (a) and the number of suggested locations (b)

deeply the possible presence of a phase transition in the knowledge acquisition process, as suggested by the previous results. Moreover, we study some general properties of the nodes' SN, such as degree distribution and clustering coefficient.



(a) Chain (b) Clust. Coeff. = 0.2 (c) Clust. Coeff. = 0.5

Fig. 8: Phase transitions for the Hit Ratio metric; #tags = 50, warm threshold = 15s

Fig. 8 presents an analysis on the phase transition for the Hit Ratio for all the three SN location organisation configurations. The results are obtained by varying the forget value and with a $T_{max} = 50$ and $warm = 15s$. Similar results can be obtained with different configurations of the parameters. Note that for the lowest forget values, all the configurations lead to a loss of information rather than an increase. This is due to the fact that the forget process is quicker than the knowledge acquisition process, so that data is dropped rapidly before it can be seen again or new data can be added. Interestingly, the Hit Ratio does not reach a value of 0. Physical Locations do not change their SNs. Hence, devices can always add some information from them, even if they drop it in a very short time. Eventually, the system reaches a stabilisation point where the Hit Ratio does not increase or decrease anymore. This behaviour is changed by increasing the forget value. In particular, for the two clustered configurations, a small increase of the forget parameter (from 50s to 75s) leads from a loss of information to a gain of knowledge. This process is achieved also for the Chain configuration, but it needs higher values of the forget parameter. This sudden change highlights the presence of a phase transition point, showing that if the data stored in memory is able to reach a given critical mass, it prompts an increase of knowledge. Under that value, even the information initially owned by each node can get lost.

With Fig. 9 we compare the properties of the individual nodes SNs with those of the Locations SN. The latter is the result of the union of all the SNs of the physical locations available in the simulation, and represents all the information that can possibly be learned in the system and, thus, it is the asymptotic SN toward which the nodes' SNs would tend in case of infinite resources. Fig. 9 shows the average CCDF of the degree distribution of the nodes' SN at the end of the simulation for all the three configurations of the SN of the locations, and for three different combinations of the forget and tag limits. In all the results, it is possible to see that the average nodes' degree distribution has the same slope as the Locations SN. Moreover, with higher values of forget and tag limits, the nodes' curves are very close to the Locations SN curve. This fact points out that the nodes are able to organise the information in their memories in such a way that it closely approximates the characteristics of the global information of the Locations SN. In particular, the nodes' SN have the presence of vertices with high degrees, even if with slightly lower probabilities than that of the global information. This is particularly relevant for the Chain configuration scenario, where devices acquire the information in the form of strings of tags. Even with the most difficult conditions for the information diffusion process, the nodes' SN self-organise in such a way that we can find hubs that allow to both bridge the description of different physical locations and determine correlations between different concepts.

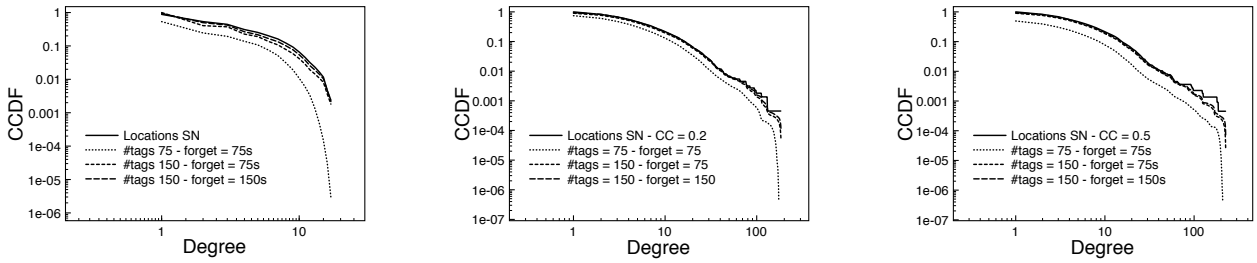
TABLE II: Average Clustering Coefficient of nodes' tag graphs at the end of the simulation

Forget		*# of tags* 75	*# of tags* 150	Locations SN CC
75	Chain	0.0195	0.0412	0.0454
	CC = 0.2	0.1632	0.2639	0.1343
	CC = 0.5	0.1557	0.3532	0.3024
150	Chain	0.0316	0.0484	0.0454
	CC = 0.2	0.2268	0.2856	0.1343
	CC = 0.5	0.3108	0.3748	0.3024

Not only hubs are replicated. In Tab. II we report the findings of the average clustering coefficient (CC) of the final nodes' SN, compared with the Locations SN. The results are shown for different values of the forget and tag limit parameters and with $warm = 25$s. Note that the clustered locations SN configurations generally lead to nodes' SN with higher CCs than that of the Locations SN. This is mainly due to the fact that the nodes SN do not hold all the vertices and edges of the Locations SN. With this still incomplete knowledge, the forget process drops more likely the edges and vertices that diminish the CC than edges and vertices that are part of clustered part of a SN. These latter components of each SN are more easily accessible from each other and, as a result, are more likely exchanged upon contact. Thus, the nodes' individual SNs organise in such a way that they learn and maintain more correlated components in memory than isolated, less connected ones. This process resembles the human brains ability to retain more closely related concepts rather than weakly associated ones. On the other hand, generally, with the most difficult initial conditions (i.e. the Chain configuration) and/or the lowest parameter settings, the knowledge acquisition process advances with more difficulties. Thus, nodes could not be able to even fetch and maintain in memory all the elements of the clustered components. As a result, the CC of the nodes' SNs is lower than that of the Locations SN.

VI. Conclusion

In this paper we have explored the use of models of human cognitive processes to design a data dissemination scheme for making users' personal devices aware of the features of the physical environment around them. Our system uses the same cognitive schemes that drive the behaviour of the human brain in acquiring knowledge about the environment and becoming aware of its features. As personal devices are proxies of their human users in the cyber world, such a direct mapping is an interesting approach to self-organisation of mobile networks.



(a) Chain (b) Clust. Coeff. = 0.2 (c) Clust. Coeff. = 0.5

Fig. 9: Mean degree distribution in the nodes' final SNs

The algorithm we have proposed is shown to be quite efficient from a number of standpoints. Under limited resources, it results in more effective dissemination of information with respect to reference solutions that do not exploit cognitive models. This result is achieved thanks to a structured information representation in memory, based on cognitive models, that allows the most relevant data to be kept in the users' devices, allowing an increase of knowledge. Moreover, the structural properties of the network of information describing physical places collected autonomously at nodes is remarkably close to what would be achieved (asymptotically) with infinite resources. Finally, a sensitiveness analysis revealed interesting properties of the system, such as the presence of phase transitions in the dissemination process, determined by the system's parameters. Interestingly, the presented results open to further investigations, such as, for example, models that can mathematically describe the properties of these transition phases, or, in a totally different domain, ways of organising the description of physical places that optimise their diffusion. Nevertheless, the set of results we have presented in the paper already provides a solid indication about the effectiveness of our approach for implementing a self-organising data dissemination scheme for making mobile nodes autonomously aware of the features of the physical environment around them.

ACKNOWLEDGMENT

This work is supported by the RECOGNITION (FP7-IST 257756) and EINS (FP7-FIRE 288021) EC projects.

REFERENCES

[1] M. Conti et al., "Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 2–21, 2012.

[2] C. Boldrini and A. Passarella, "Data dissemination in opportunistic networks," in *Mobile Ad Hoc Networking: Cutting Edge Directions, Second Edition*, S. Basagni et al., Ed. Wiley, 2013, pp. 453–490.

[3] G. Gigerenzer, "Why heuristics work," *Perspectives on Psychological Science*, vol. 3, no. 1, pp. 20–29, 2008.

[4] L. J. Schooler, R. Hertwig *et al.*, "How forgetting aids heuristic inference," *Psychological review*, vol. 112, no. 3, pp. 610–627, 2005.

[5] M. Conti, M. Mordacchini, and A. Passarella, "Data dissemination in opportunistic networks using cognitive heuristics," in *Proc. of AOC 2011*. IEEE, 2011, pp. 1–6.

[6] R. Bruno, M. Conti, M. Mordacchini, and A. Passarella, "An analytical model for content dissemination in opportunistic networks using cognitive heuristics," in *Proc. of MSWIM 2012*. ACM, 2012, pp. 61–68.

[7] L. Valerio, M. Conti, E. Pagani, and A. Passarella, "Autonomic cognitive-based data dissemination in opportunistic networks," in *Proc. of IEEE WOWMOM 2013*, 2013, pp. 1–10.

[8] M. Conti, M. Mordacchini, A. Passarella, and L. Rozanova, "A semantic-based algorithm for data dissemination in opportunistic networks," in *Proc. of IWSOS 2013*. Springer, 2013.

[9] M. Mordacchini, L. Valerio, M. Conti, and A. Passarella, "A cognitive-based solution for semantic knowledge and content dissemination in opportunistic networks," in *Proc. of AOC 2013*. IEEE, 2013.

[10] M. J. Egenhofer and D. M. Mark, "Naive Geography," in *Spatial Information Theory A Theoretical Basis for GIS*, 1995, pp. 1–15.

[11] Y. Tuan, *Space and place: The perspective of experience*. University of Minnesota Press, 2001.

[12] M. Raubal, H. J. Miller, and S. Bridwell, "User-Centred Time Geography for Location-Based Services," *Geografiska Annaler: Series B, Human Geography*, vol. 86, no. 4, pp. 245–265, 2004.

[13] C. Clark and D. L. Uzzell, "The Affordances of the Home, Neighbourhood, School and Town Centre for Adolescents," *Journal of Environmental Psychology*, vol. 22, no. 1-2, pp. 95–108, 2002.

[14] J. J. Gibson, *The Ecological Approach to Visual Perception*. Lawrence Erlbaum, 1986.

[15] T. Jordan, M. Raubal, B. Gartrell, and M. Egenhofer, "An Affordance-Based Model of Place in GIS," *SDH 98*, vol. 98, pp. 98–109, 1998.

[16] B. Bennett and P. Agarwal, "Semantic categories underlying the meaning of 'place'," *Proc. of COSIT 2007*, pp. 78–95, 2007.

[17] M. Zook and M. Graham, "Mapping digiplace: geocoded internet data and the representation of place," *Environment and planning*, vol. 34, no. 3, p. 466, 2007.

[18] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proceedings of ACM SIGIR 2007*. ACM, 2007, pp. 103–110.

[19] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using flickr tags to describe city cores," *Journal of Spatial Information Science*, no. 1, pp. 21–48, 2012.

[20] J. Cranshaw et al., "The livehoods project: Utilizing social media to understand the dynamics of a city," in *Proc. of ICWSM 2012*, 2012.

[21] R. Stedman, "Toward a social psychology of place predicting behavior from place-based cognitions, attitude, and identity," *Environment and behavior*, vol. 34, no. 5, pp. 561–581, 2002.

[22] G. B. Colombo, M. J. Chorley, V. Tanasescu, S. M. Allen, C. B. Jones, and R. M. Whitaker, "Will you like this place? a tag-based place representation approach," in *Proc. of PerMoby 2013, San Diego*, 2013.

[23] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, Nov. 2007.

[24] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198–208, Apr. 2006.

[25] E. Quintarelli, "Folksonomies: power to the people," in *Proc. of ISKOI 2005*, 2005.

[26] R. Sinha, "Tagging from personal to social - WWW 06 keynote," Tech. Rep., 2006.

[27] B. Gawronski and B. K. Payne, *Handbook of implicit social cognition: Measurement, theory, and applications*. The Guilford Press, 2010.

[28] R. S. Wyer Jr, "Principles of mental representation," *Social psychology: Handbook of basic principles*, vol. 2, pp. 285–307, 2007.

[29] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[30] S. Deyne and G. Storms, "Word associations: Network and semantic properties," *Behavior Research Methods*, vol. 40, no. 1, pp. 213–231, 2008.

[31] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical review E*, vol. 65, no. 2, p. 026107, 2002.

[32] A. Vahdat, D. Becker *et al.*, "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke University, Tech. Rep., 2000.