Efficient Social-aware Content Placement in Opportunistic Networks

Panagiotis Pantazopoulos[†]   Ioannis Stavrakakis[†]   Andrea Passarella[‡]   Marco Conti[‡]

[†]Department of Informatics and Telecommunications
National & Kapodistrian University of Athens
Ilissia, 157 84 Athens, Greece
Email: {ppantaz, ioannis}@di.uoa.gr

[‡]Institute for Informatics and Telematics - CNR
via G. Moruzzi, 1
56124 Pisa, Italy
Email: {a.passarella, marco.conti}@iit.cnr.it

Abstract—As content provisioning becomes the driving application of today's (opportunistic) networking environments and the User Generated Content explodes, the problem of devising scalable approaches to placing it optimally within a networking structure becomes more important and challenging. Since the well-known *k-median* optimization problem that is typically formulated to address it requires global topology and demand information, different approaches are sought for. The latter is the focus of this paper that aims at exploiting social structures, present in emerging networking environments, in order to devise a *scalable* approach to the optimal or near-optimal content placement. A new metric that captures the node's social significance or potential for helping establish paths between nodes is introduced and serves as the basis for creating a small scale network sub-graph over which the small-scale content placement problem is solved sequentially until the optimal or near-optimal location is identified. The trade-off between the sub-graph's size and the degree of convergence to the optimal solution is studied through simulations on E-R and B-A random graphs and the effectiveness of the proposed approach is demonstrated.

I. Introduction

Opportunistic networks [1] are self-organizing wireless mobile networks formed by user devices, without requiring any pre-existing network infrastructure. Communications in this environment has been typically supported through Mobile Ad hoc Networks (MANETs). However, MANET solutions work only if a rather stable topology can be established among the nodes, which is often not the case in the presence of users' mobility. Opportunistic networks support communication among nodes even when no stable multi-hop paths between communication endpoints can be established. In an opportunistic network a node carrying a content addressed to a given destination evaluates if any other node it comes in direct contact is "better suited" than itself to bring the content to the destination. In other words, each contact is opportunistically exploited to bring the content closer and closer to the destination.

While research on opportunistic networks has initially focused on forwarding issues, recently attention has also been paid to *data dissemination* problems. Opportunistic networks are formed by user devices, which are increasingly capable of generating complex multimedia *User Generated Content*. Devising solutions for sharing content through decentralized data dissemination techniques is clearly a challenging and important issue. On the other hand, *social awareness* (i.e., information about the social relationships and interactions among the users) is increasingly perceived as a key contextual information to design networking solutions for opportunistic networks. With respect to the data dissemination problem, for example, solutions have been proposed that leverage information about how users group in social communities to design efficient content replication schemes [2], [3].

This paper falls in this stream of research, and studies the problem of optimal placement of a given piece of content in an opportunistic network. First, we discuss that contact patterns among users naturally hint to a graph representation of the network, where a link exists among nodes if they are "enough frequently" in touch with each other. Given this representation, existing literature approaches the problem of content optimal placement as an instance of the facility location problem, and identifies heuristics in order to deal with the fact that this optimization problem is -for the general case- NP-hard.

In this paper we propose a social-aware heuristic to find the content's optimal location. Intuitively, content is iteratively migrated to nodes that are increasingly "central" to the overall network, i.e. nodes such that the average cost of accessing the content from any other interested node is increasingly lower. We analyse the performance of the proposed heuristic, which works with local information, studying its degree of approximation of the optimal centralized solution on two types of graphs representing different user contact patterns, i.e. the Erdős-Rényi (E-R) [4] and the Barabási-Albert (B-A) [5] graph model. Interestingly enough, we find that the heuristic performs much better -and close to the optimal- under the B-A model, which is known to represent more realistically

than the E-R model user social interactions and contact patterns.

II. A GRAPH-ORIENTED APPROACH

In this section we firstly discuss why opportunistic networks can be represented with random graphs. Then, we provide an overview of the resulting formulation of the optimal content placement problem, and finally present the rationale behind the proposed social-aware heuristic.

A. A Social-oriented graph representation of opportunistic networks

In opportunistic networks communication links between nodes appear and disappear dynamically over time. Nevertheless, it is possible to represent an opportunistic network with a logical graph structure, by adding an edge between two nodes if the frequency of encounters between the two nodes or the cumulative time they have been in direct contact (over a given time window), exceeds a pre-defined threshold. This representation has been already proposed in the opportunistic networking literature, and has been exploited to design distributed community detection algorithms [6].

Here, we build on this way of representing an opportunistic network to study the optimal content placement problem. Note that a multi-hop path on such a graph does not imply a simultaneous multi-hop path between two nodes as in conventional MANETs. Instead, a multi-hop path means that it is possible to move content from one path end-point to the other by opportunistically exploiting contacts between the nodes in the path; such contacts are not necessarily simultaneous. Therefore, a multi-hop path in such a graph has to be seen as a space-time multi-hop path [7].

Note that this graph representation implicitly includes social information about the nodes. Recall that an edge represents the fact that two nodes have been in touch for at least some amount of time or with at least some frequency. Therefore, it means that those nodes are co-located with each other often enough, and thus they are likely to share some social interest. This representation does not distinguish between the case in which nodes get in touch because the users actually share a social relationship, and the case in which users do not know each other but happen to be often physically co-located. However, from the standpoint of communication protocols, this does not make a huge difference. In both circumstances, it is a fact that the two nodes are often in touch with each other due to some feature of their users' social behaviour, and this direct communication link can be opportunistically exploited to optimize the networking operations (optimally placing content in our case).

In view of this social-oriented graph representation of opportunistic networks, the problem of optimal content placement can be seen as how to optimally place a given content on a graph. Traditionally, this problem has been tackled as an instance of the facility location problem [8], one of the most studied problems in operations research, which considers the optimal placement of some facilities serving the demands of a set of clients. The main question is to determine, in the general case, the number and locations of the facility set elements, in a way that the cumulative cost of serving all clients' demands, is minimized.

B. The facility location problem formulation

Having adopted the above introduced graph representation, we aim to identify the network nodes of a given topology that can host some content, so that the latter is accessed by the network users residing at the various network nodes with the minimum average cost.

The network topology is represented by an undirected connected graph $G(V, E)$, where V is the set of nodes and E is the set of edges, between them. Without loss of generality (see discussion later) we assume that all links have a unit of weight and thus the minimum cost path $d(n, m)$ between nodes n and m, corresponds to the shortest path between n and m. The demand for the content, generated by a node n, is denoted by w_n. When there are k replicas of the content available, the problem of their optimal placement can be formulated as the classical *k-median* problem that determines the set $\mathcal{F}$ of k host-nodes ($|\mathcal{F}| = k$), that can host the content, so that the following average content serving cost is minimized:

$$Cost(\mathcal{F}) = \sum_{n \in \mathcal{V}} w_n \cdot min_{x_j \in \mathcal{F}} \{d(x_j, n)\} \qquad (1)$$

where $min\{d(x_j, n)\}$ is the smallest distance among all, from each host-node $x_j \in \mathcal{F}$ to node n. In this paper we focus on the case in which only one copy of the content is available; this case may capture well the spontaneous generation of the aforementioned *User Generated Content* (UGC) that is of relatively low demand (thus its replication would probably not be sensible or affordable) and seeks to find its optimal placement within a network. Thus, in this case $|\mathcal{F}| = 1$, yielding a *1-median* formulation with the content located at node $k \in V$ and a cost to be minimized given by:

$$Cost(\mathcal{F}) = \sum_{n \in \mathcal{V}} w_n \cdot d(k, n) \qquad (2)$$

Generally speaking, optimization problems like the above ones are NP-hard (even when the service cost is considered to be the Euclidean distance from a node to the nearest content's host [9]), requiring global information about the network topology and generated demand load. Thus, the main bulk of relevant theoretical work, so far, is in the field of approximation algorithms, where various techniques have been applied (e.g. rounding to LP relaxation [10] or primal-dual methods [11]).

C. Social based small-scale approach to overcome limitations

Herein, we propose and evaluate an algorithm which incorporates a social dimension in the problem of content's optimal placement by exploiting information stemming from the nodes' social attributes. In an opportunistic networking environment nodes can have fairly diverse characteristics with respect to the kind and frequency of social communication links with other nodes. From a networking standpoint, this means they can have different roles as intermediaries, as, for example, nodes having a higher number of links might be key in helping establish links between other nodes. Such nodes could also probably be major players in the solution of the *1-median* problem. The main objective of this paper is to identify them, and use them to define a thinned-out network topology yielding a smaller-scale optimization problem with respect to the one involving all the nodes of the graph.

To this end, an innovative social metric is introduced, referred to as *conditional betweenness centrality* (to be defined later) and a thinned-out subnetwork topology is identified by considering the nodes with the highest values of this metric. It turns out that this metric possesses two important features: first, it yields subnetworks of nodes that can be constructed with local information and are contained within a generally confined locality, which is important for the implementation and scalability of the proposed solution; second, it captures implicitly the content demands of the network nodes not included in the thinned-out subnetwork, which should help reduce the approximation error.

The rest of the paper is organized as follows. In Section III we give an overview of the related work and in the next one we introduce highlights of our approach. In Section V, we introduce a novel social-inspired metric that will be used, later on, in our algorithm which is analyzed in Section VI. The Section VII contains our experimental results and comments on the performance of the proposed solution. Finally, in Section VIII we conclude the paper and discuss possible extensions.

III. Related work

Data dissemination problems (of which content placement is an example) are starting being investigated in opportunistic networks. The work in [2] proposes ContentPlace, which is a social-oriented framework for data dissemination. ContentPlace builds on the work in [6] and assumes that nodes can be aware of the social communities they belong to. Using a general utility-based optimization framework, ContentPlace defines distributed algorithms for nodes to select which content to locally replicate, out of what is available on encountered nodes. These algorithms take into consideration the estimated distribution of content in the network, and the interests of the users with respect to content. A similar approach is taken in [3] where

it is shown that mobility and cooperative content replication strategies can help bridge social groups. The work in [12], which also relies on [6] for community detection, defines a social-oriented pub/sub overlay for opportunistic networks. One broker per community is selected, as the node being more "central" (from a social standpoint) in the community. A broker overlay is maintained, and publications are disseminated to all brokers, which then decide to which subscribers they should be forwarded. The work presented in this paper provides some solid grounds to the choice of selecting "central" nodes for content placement.

As already mentioned, there is a great amount of work dealing with the facility location problem. Typically centralized approximation algorithms using various techniques (like rounding fractional solutions to linear programming relaxations [13]) have been employed to solve the problem, associated with large scale optimization. Heuristic methods like local search [14] have also been used to approximate the optimal solution in polynomial time. When considering the limitations of dynamic networking environments, the above facility location problem becomes more challenging and has received renewed attention [15]. Ideas like iterative negotiation between clients and service instances [16] or a plethora of application layer solutions have been proposed. An identical to the *1-median* problem has been studied in the work of Oikonomou et al. [17], where a single service facility moves one hop at a time, following a cost decreasing path, when some service-movement criteria are fulfilled.

Our work, which has the above focus, was motivated by the upcoming paper of G.Smaragdakis et al. [18]. In there, the placement problem has been solved in a distributed manner by using some limited scope centralized approach within an area of *r-hops* from the location of each service and mapping the demand coming from the outer nodes on the ring of the *r-hop* area. Still, the content placement problem for a social network or even the incorporation of some social dimension to its solution, has not been considered.

IV. Framework of proposed solution

In this section we highlight the key points of our approach so as to bring out some important details. Social information is being exploited here to determine the optimal physical location of the node to host some content by solving iteratively a spatially restricted, low-cost *1-median* problem (instead of a costly, global one). Consequently, the proposed algorithm can provide for easy adaptation to dynamic environments due to its local-info requiring approach. As approximations are involved, the proposed solution may not always lead to the identification of the optimal location; nevertheless, it is shown that the divergence error is not substantial.

Assuming full topology knowledge, over a limited region around some node currently hosting the content, we propose a social based criterion, for selecting a number

of fairly neighbouring nodes (forming a subgraph) to take part in a small-scale, local solution. This, may be derived by employing any well-studied centralized approach (e.g. approximation algorithm, heuristics). The selected nodes (a percentage of total number) are the ones, having the top values of an innovative metric, inspired by social network studies, that plays a twofold key role. First, it captures a node's significance, regarding its capability to transport content efficiently. Second, it captures the contribution (to the *1-median* solution) of incoming demand from the rest of the network nodes, not included in the above subgraph. Actually, we show that the demand, following a uniform model, can be deduced from this metric which is a measure of the role of every node in the information flow, towards the one having the content.

After solving the small-scale optimization problem described above, the node within the subgraph considered, that minimizes the content provision cost (if hosting the content) is identified. Then, the content is assumed to be placed in this node[1] and the new subgraph is determined around this node and the new small-scale optimization problem is solved. This iterative procedure repeats until no further movement of the content is possible; that is, the node with the minimum cost turns out to be the currently hosting node. Notice that the content moves according to the outcome of the optimization, on a cost-decreasing path, trying to reach the optimal location.

V. Conditional Betweenness Centrality

The solution proposed herein relies heavily upon the criterion based on which the subgraph, around the current node having the content, is picked in order to solve (there) the small-scale optimization problem. This criterion (referred to as CBC criterion hereafter) involves an innovative, to the best of our knowledge, metric coming from social networks analysis, that captures network traffic towards a specific node (the node, each time, hosting the content).

Centrality measures (or indices) are widely used in social network analysis, since Freeman's late '70s influential articles [20], [21]. Actually, they are used as graph-theoretic tools in order to explain social phenomena. These indices, defined either on the nodes or edges of a graph, are usually based on geodesic paths that link members of a social network and aim to provide a measure of importance of their social position, under the assumption that importance is equally divided among all shortest paths of each pair. Different measures have been introduced to capture a variation of a node's importance, like ability to reach numerous nodes via relative short paths or popularity among others [21].

Betweenness centrality, one of the most frequently used metric, is a measure of the extent to which a node lies on the (shortest) paths linking other nodes. Let σ_{st} denote the number of shortest paths between any two nodes s and t in a connected graph $G = (V, E)$. If $\sigma_{st}(u)$ is the number of shortest paths passing through the node $u \in V$, then the *betweenness centrality* index of node u is given by (3).

$$BC(u) = \sum_{s=1}^{|V|} \sum_{t=1}^{s-1} \frac{\sigma_{st}(u)}{\sigma_{st}} \qquad (3)$$

$BC(u)$ captures a node's u ability to control or help establish paths between pairs of nodes; this is an average over all network pairs (see (3)). When the content is hosted by a specific node t, the traffic flow of relevance -that shapes the resulting cost of content provisioning from that node- is the one between all node pairs (x, t) $\forall x \in V$ for the fixed node t, and not all possible pairs, as in (3). Consequently, if we were to select a subgraph of nodes to solve a small-scale optimization problem, it would make sense to include the nodes that stand between the "most" paths linking the network nodes to the specific one hosting the content; the presence of such nodes would reflect somewhat the fact that relatively high demand (that shapes the resulting cost) is coming through such nodes. The *conditional betweenness centrality* (CBC) index defined by (4), can be used as a criterion for constructing the network subgraph (assuming $\sigma_{st}(s) = 0$):

$$CBC(u; t) = \sum_{s \in V, u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} \qquad (4)$$

A possible implementation of the calculation of the CBC index could take place by having the nodes monitor the packets routed through them. Assuming that routing takes place along shortest paths on the considered space-time graph and that packets carry with them the already traveled path, the quantities involved in (4) can be computed by each node t and the $CBC(u; t)$ values for all nodes u can be locally (at node t) available.

Based on the definition in (4), it is expected that nodes u which lie in between t and some dense networks region will exhibit higher CBC values, due to the large number of shortest paths passing through u coming from that region and leading to node t. The closer a node u is to t, the greater the number of nodes that reach t through it and thus, the larger the CBC values for such u nodes. Consequently -and as it is also discussed later in conjunction with Fig.1 and Table I- the subgraph containing the nodes with the higher CBC values is expected to be confined with a locality around t.

VI. Analysis of 1-Median social-aware approach

In this section we present the social-aware approach to the content placement problem. Our solution's pseudocode is depicted in the Algorithm 1, where the cost of placing

---

[1]Although out of scope of this work, there are various component migration mechanisms at application layer, as well. They are provided by some distributed architectures and are able of moving both functionality and data from one host to another, in a transparent way to the user [19]. This implies that practical software solutions may use an idea like the proposed one, without prohibitive cost.

the content at node k of graph $G(V, E)$ that represents the entire topology, is denoted by $C(k)$. The i-th iteration's subgraph around the host-node n, where the small-scale optimization takes place, is denoted by G_n^i. Finally, we use the variables $C_{current}$ and C_{next} to hold the cost of current and next step solution, respectively.

Algorithm 1 Social-aware *1-median* in *G(V,E)*

---

 1. *choose randomly node s*
 2. *place CONTENT in s*
 3. $C_{current} \Leftarrow \infty$
 4. $\forall\ u \in G$ *compute* $CBC(u; s)$
 5. *let* G_s^o *be* $\alpha\%$ *of G nodes with top CBC values*
 6. *1-median solution in* $G_s^o\ \rightarrow\ node\ Host$
 7. $C_{next} \Leftarrow C(Host)$
 8. **while** $C_{next} < C_{current}$ **do**
 9. *move CONTENT to Host*
10. $C_{current} \Leftarrow C_{next}$
11. $\forall\ u \in G$ *compute* $CBC(u; Host)$
12. *let* G_{Host}^i *be* $\alpha\%$ *of G nodes with top CBC values*
13. *1-median solution in* $G_{Host}^i\ \rightarrow\ node\ NewHost$
14. $Host \Leftarrow\ NewHost$
15. $C_{next} \Leftarrow C(NewHost)$
16. **end while**

A. Detailed description

The proposed algorithm begins with the random choice of node s where the content is initially placed. To ensure the first hop of our cost-decreasing path, we assign infinite value to the cost of that placement in order to fulfil the cost condition (line 8) at least once. Next, the computation[2] of $CBC(u; s)$ metric takes place for every node u in G. Nodes having the top CBC values comprise subgraph G_s^o, the size of which is determined by the percentage $\alpha\%$ of total number of network nodes. The following step (line 6) is to solve the *1-median* problem on the subgraph G_s^o, consisted of $\alpha \cdot V$ nodes. Note that the CBC index cannot be defined for the host-node which is not included in the solution space (i.e. nodes that are possible solutions for the *1-median* problem).

Any centralized technique may be used to solve the small-scale optimization problem (on each iteration's selected subgraph). One can seek for the best heuristic method available (successively better algorithms have been designed during the last few years [24]), so as to achieve maximum scalability. For simplicity reasons we have solved it using enumeration [8], even if there are other more cost-effective methods. Besides, this implementation choice

does not affect the degree of convergence of our solution to the optimal one.

The optimization's outcome is the location of node $Host$, with which we achieve minimum service cost[3], called $C(Host)$, among the nodes of the current subgraph. We assign the value of this cost to the variable C_{next} and test whether the cost-decreasing condition ($C_{next} < C_{current}$) is fulfilled. In case it is true, the content is moved to node $Host$ and the same steps are followed thereafter[4]. As long as the aforementioned condition remains true, the above process is executed iteratively and the content migrates from node to node, trying to reach the (globally) lowest-cost one, to be hosted at.
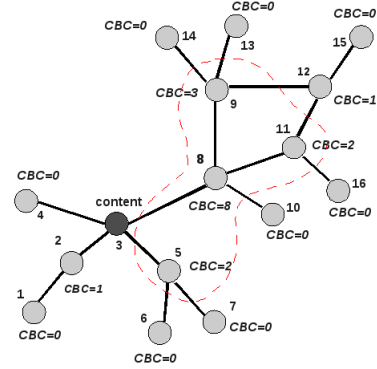

Fig. 1. Encompassing dashed line marks those nodes singled out by $CBC(u; 3)$ criterion ($\forall u \in G$).

In the sequel, we discuss how the CBC criterion applied here, results in subgraphs that manage to "forward" the content to more cost-effective regions. Assume we have the network topology of Fig. 1, where the content has been placed at host-node '3'. When applying the CBC criterion, we pick those nodes that possess, socially significant, intermediate places between the content's host-node and a great number of other nodes. The nodes selected[5], included in the dashed line of Fig. 1, are fairly neighbouring to the host, though not forming some sort of circular area, around it. Rather, they deviate from exact surrounding locality by stretching in a certain direction (towards nodes '8', '9' and '11'). This deviation is due to the presence of some nodes with significant social role, regarding their participation in numerous information pathways (that correspond to shortest paths) from the rest of the nodes to the content's host. So, the solution space of our *1-median* problem is practically moved by this selection, towards directions/areas "populated" by nodes highly capable of transporting content, efficiently. This CBC characteristic of identifying a socially significant

---

[2]For our simulation's needs, this involves solving the all-pairs shortest path problem. Common algorithms, like Floyd-Warshall [22], may need even $\Theta(|V|^3)$ time to solve, on a $G(V, E)$ graph. Hence, for CBC computation we properly modified a new, scalable algorithm [23] for *betweenness centrality*, with runtime $\mathcal{O}(|V||E|)$. The cost introduced is low, as the length and number of all shortest paths from a given source to every other node, needed for our computation, is determined in $\mathcal{O}(|E|)$ time [23].

[3]In case of multiple minimum-cost solutions within the G^i nodes, we choose randomly one of them.

[4]The percentage α of total number of nodes, that forms the G^i subgraph, remains the same until our algorithm's completion.

[5]In Fig.1, let the percentage α be of such a magnitude that results in picking the nodes of the 3-rd top CBC level.

"direction" within the network is valid under any demand model but can be useful and exploited only under the uniform demand hypothesis. In the latter case, the nodes with high CBC values are those that have "behind" them large portions of the network and thus (under uniform load) of the demand traffic as well. When non-uniform demand is involved these directions may be misleading since the existence of a great number of nodes connected to a (thus) significant one, does not correspond to great demand load stemming from them.

B. On the convergence of proposed solution

In this paragraph we examine whether the cost-decreasing criterion is sufficient enough to make the content's migration mechanism halt at some host-node. We sketch a proof for the convergence below:

Theorem 1: Algorithm 1 converges to a solution after a finite number of steps.

Proof: We distinguish three cases that describe *all* the different ways of content's migration, according to Algorithm 1. For each one we show that convergence is achieved.

α. The content reaches node n where $C(n) < C(m)$, $\forall m \in G_n^1$ which is the first subgraph formed by the CBC criterion. Thus, the algorithm terminates there.

β. The content moves in a loop-free fashion amongst the nodes of G. This means that the content, following a cost-decreasing path, may be hosted at some node u at most one time. Since the solution space is finite, the content's migration will stop at the k-th iteration, where $k \in [1, V]$. Thus, the algorithm terminates there.

γ. The content moves in an infinite loop fashion. The situation involves the consecutive migration of the content, from one host to another, visiting the same nodes infinite times. This scenario is shown below to contradict the cost-decreasing criterion, so our proof is completed. Suppose that the content, moved by Algorithm 1, reaches some node $b \in G$, twice. Right after its first placement at b, we solve the *1-median* in the subgraph $G_b^{[1]}$ (index $[k]$ denotes the k-th time when the content is placed at b) that is formed by the application of $CBC(u; b)$ criterion, $\forall u \in G$. Let the corresponding cost be $C_b^{[1]}$. When the content returns to b, given that the network topology remains the same, the deterministic CBC criterion of (4) singles out the same subgraph (around b) with the one of the first visit, so we have that $G_b^{[1]} = G_b^{[2]}$, implying for the costs that $C_b^{[1]} = C_b^{[2]}$, which contradicts the cost-decreasing condition of Algorithm 1. Thus, the content cannot be moved any farther than node b and consequently, we conclude that once the content reaches some node for the second time, Algorithm 1 terminates. ∎

VII. Simulation Results

We have implemented the algorithm described in section VI in our simulation environment. Experiments, presented in the next paragraphs, have been carried out on well-studied random graph models, in order to evaluate its performance.

A. Simulation settings

In our set-up, we generate both Erdős-Rényi [4] and Barabási-Albert [5] random graphs of N=100 nodes. For the first ones, each edge is included in the graph with probability p=0.4, while the second ones are generated by starting from an initial network of m_0=10 nodes and each newly added one connects to m=2 other nodes according to the well known preferential attachment rule.

For any chosen configuration of the parameters, we replicated simulation runs 10 times to achieve statistical significance. Specifically, results presented hereafter are the averages over the simulation runs. Note that the graphs are re-created at the beginning of each simulation run. This allows us to reproduce the key statistical properties of the chosen graph models. It is easy to see that, as we re-create the graphs for each simulation runs, over a sufficiently large number of replicas, the probability of any given link to exist approximates the theoretical value of the target graph model. For example, in the case of E-R graphs, it is easy to show that, as the number of replicas increase, the overall probability of any specific edge among any two nodes exist with probability equal to p. Finally, note that this also results in meaningful test cases from a social perspective. In the case of E-R graphs, each pair of users have the same probability of getting in touch with each other, which corresponds to a completely uniform opportunistic environment. This is not a particularly realistic model, but is definitely a reference case in the literature, and is thus included as a benchmark. The case of the B-A model reproduces a more realistic scenario, in which the number of social relationships of a given node is distributed according to a power law.

B. Experiments on random graphs

First, we measure the locality of our solution by computing the mean percentage of the nodes in the current subgraph $G^i \subseteq G$, which are *not* within a radius of two hops away from content. Having these nodes in a small radius around content's location, is desirable to keep the communication cost low. We call this mean percentage, that comes out of 10 runs, *mismatch* and present its values in Table I. Obviously, such an attribute is no longer meaningful when considering percentage α close to 1. The results in Table I suggest, as expected, that the CBC criterion picks fairly neighbouring nodes around the current content's location, showing a locality attribute of the proposed algorithm.

We now present the results illustrating the degree of convergence of our social-aware solution to the optimal one, derived assuming global topology and demand information. We run our algorithm on both E-R and B-A graphs assuming a unit of demand from all nodes

TABLE I
MISMATCH FOR VARIOUS SIZES OF SUBGRAPH G^i

α of G nodes (%)	E-R mismatch (%)	B-A mismatch(%)
10	23.7	1.5
20	36.0	3.5
30	43.4	11.3
40	56.5	19.9
50	62.7	28.7
60	68.2	31.0
70	71.3	33.6

$(w_n = 1, \forall\ n \in G)$. We measure the mean ratio of our algorithm's cost $C_{algorithm}$ to the cost of the optimal solution $C_{optimal}$ and call it $\beta_{algorithm}$:

$$\beta_{algorithm}(\alpha) = E[\ \frac{C_{algorithm}}{C_{optimal}}\ ] \qquad (5)$$

Fig. 2 depicts the mean value of the normalized cost ratio β, for various values of the percentage of G nodes participating in the solution. The social-aware approach
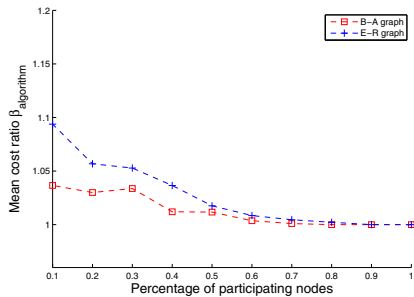

Fig. 2. Convergence of proposed solution to optimal, against the percentage of participating nodes

seems good convergence to the optimal, yielding a low-cost solution. Clearly, the larger the percentage of nodes included, the smaller the induced error. The better performance of our algorithm for B-A than E-R graphs is due to the preferential attachment characteristic of the first ones [5], which captures one key (social) feature of real-world networks and leads to the appearance of highly connected nodes, capable of serving (as host-nodes) the demands at fairly low cost.

Finally, we compare the cost induced by the derived social-aware solution to that induced by placing the content in a randomly selected node; the latter would amount to serving the demands from the node that, e.g., generated the content, without paying the price of taking consecutive steps towards less costly positions. We measure the mean ratio between the cost of the random policy and the cost of the optimal (global) solution, defined likewise (5). This ratio is called β_{random}:

$$\beta_{random} = E[\ \frac{C_{random}}{C_{optimal}}\ ] \qquad (6)$$

In order to compare the two approaches we choose the maximum cost ratio $\beta_{algorithm}$, shown in Fig. 2, corre-

sponding to size of subgraph $\alpha = 10\%$ of total G nodes. The results below (Table II) show that our algorithm outperforms -by far- the random policy, even when considering the most costly scenario, from Fig. 2.

TABLE II
SOCIO-AWARE *1-median* VS SIMPLE RANDOM PLACEMENT

E-R graphs		B-A graphs	
max. $\beta_{algorithm}$	β_{random}	max. $\beta_{algorithm}$	β_{random}
1.0938	1.2736	1.0366	1.4953

C. Demand mapping

In this paragraph we talk about a variation of the proposed solution, worthy of remark. The idea here, is to drop the initial assumption that each node $k \in G^i$ takes part in the *1-median* solution with the same demand load $(w_k = 1, \forall k \in G^i)$ and apply some sort of mapping between the CBC values of the nodes $k \in G^i$ and the incoming demand load, these nodes may receive from the nodes not included in G^i. As stressed above, the CBC metric can be interpreted as a node's capability of information flow control towards the content's host-node, by occupying a significant, intermediate among others, position. Consequently, the *higher* the CBC values for a node, the *greater* the amount of information flows through it (heading towards the current host-node). But, the fact that information flow reflects the aforementioned demand load, motivated us to introduce a new demand mapping for the small-scale optimization problem in order to improve our algorithm's performance.

In more detail, a demand model is employed to capture more properly the fact that nodes with high CBC values have increased contribution to the *1-median* solution. We implemented the following mapping from the CBC values to the demand w_k values, given by the equation[6] below:

$$w_k = CBC(k; Host) + 1, \quad k \in G^i \qquad (7)$$

The results derived from our experiments show that the presumable idea of demand mapping affects the performance of our algorithm, especially for a range of relative low values of the α percentage. When the number of nodes participating in the small-scale optimization is very small, namely $\alpha \leq 0.1$ (see Fig. 3), the demand mapping results in current algorithm outperforming the previous no-mapping approach. This fact suggests the use of the CBC values as "weights" to capture the contribution of distant demand load to our restricted optimization problem.

When considering the case of high percentage α, demand mapping does not reflect our initial arguments of taking into account the load of a significant number of distant nodes, not included in the small-sized subgraph G^i. On the contrary, the number of nodes not included

[6]Even if a node has $CBC = 0$, its demand may be different than zero. This is why we need to add one to all CBC values.
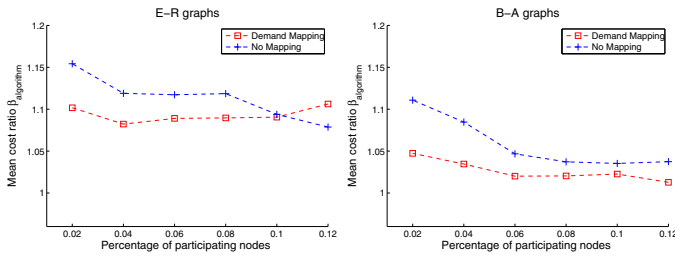
Fig. 3. Performance of very small-scale solution for E-R and B-A graphs, under mapping and no-mapping demand approach

in G^i is, in this case, low and hence the mapping of a small number of nodes' demand on the great number of G^i ones, is to no avail. Moreover the solution derived when demand mapping is applied, is affected by the CBC values which contribute, as demand load, to the cost (to be minimized) of (2). Consequently, for very high α percentages, asymptotically reaching 100%, we end up solving a completely different problem to the one with no mapping, where $w_k = 1 \forall k \in G^i$.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper we have developed a scalable and efficient approach for determining the optimal or near-optimal placement of (User Generated) content in an opportunistic networking environment, exploiting available social structures. The presented approach amounts to solving the *1-median* problem in a scalable manner.

We employed an already existing in literature graph representation of the opportunistic networks and proposed a heuristic method in each iteration of which, a limited horizon approach is applied to get a local solution based on nodes' social significance. This attribute stems from the extent to which they hold an intermediate position between the current content's host and the rest of network nodes. The out-of-horizon demand is inferred by the value of an innovative sociometric index, introduced to quantify this significance and accordingly pick the nodes to take part in the local solution. Having exploited the above social-inspired index, there is no need to apply any mapping mechanism for capturing this demand load. The extra communication and computation cost (e.g. mapping errors) occurring by possible application of such a mechanism, is (here) avoided.

Our solution is not guaranteed to always reach the optimal, but simulation shows that it achieves satisfying convergence, especially when applied on graphs incorporating with some real-world social characteristics (such as B-A random graphs). We plan to extend this work to a general solution, since the model studied in this paper is constrained according to a uniform demand hypothesis. If the latter is relaxed, an augmentation of the definition of the introduced CBC metric will be necessary to account for it. In addition, other attributes and characteristics such as correlated load demands should be considered in order to better model real-world social characteristics of user nodes communicating opportunistically.

REFERENCES

[1] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *IEEE Communications Magazine*, vol. 44, no. 11, p. 134, 2006.
[2] C. Boldrini, M. Conti, and A. Passarella, "ContentPlace: social-aware data dissemination in opportunistic networks," in *Proceedings of the 11th international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM New York, NY, USA, 2008, pp. 203–210.
[3] E. Jaho and I. Stavrakakis, "Joint interest- and locality-aware content dissemination in social networks," in *IFIP/IEEE WONS*, Snowbird, Utah, USA, February 2-4, 2009.
[4] P. Erdős and A. Rényi, "On random graphs. I," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.
[5] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct 1999.
[6] P. Hui, E. Yoneki, S. Y. Chan, and J. Crowcroft, "Distributed community detection in delay tolerant networks," in *MobiArch '07*. New York, NY, USA: ACM, 2007, pp. 1–8.
[7] V. Borrel, M. H. Ammar, and E. W. Zegura, "Understanding the wireless and mobile network space: a routing-centered classification," in *Challenged Networks*, 2007, pp. 11–18.
[8] P. Mirchandani and R.Francis, *Discrete location theory*, John Wiley and Sons, 1990.
[9] N. Megiddo and K. J. Supowit, "On the complexity of some common geometric location problems," *SIAM Journal on Computing*, vol. 13, no. 1, pp. 182–196, 1984.
[10] I. Baev, R. Rajaraman, and C. Swamy, "Approximation algorithms for data placement problems," *SIAM Journal on Computing*, vol. 38, no. 4, pp. 1411–1429, 2008.
[11] T. Moscibroda and R. Wattenhofer, "Facility location: distributed approximation," in *PODC '05*, 2005, pp. 108–117.
[12] E. Yoneki, P. Hui, S. Chan, and J. Crowcroft, "A socio-aware overlay for publish/subscribe communication in delay tolerant networks," in *MSWiM '07*, Chania, Greece, 2007.
[13] D. B. Shmoys, E. Tardos, and K. Aardal, "Approximation algorithms for facility location problems (extended abstract)," in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1998, pp. 265–274.
[14] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit, "Local search heuristic for k-median and facility location problems," in *STOC '01*. ACM, 2001, pp. 21–29.
[15] G. Wittenburg and J. Schiller, "A survey of current directions in service placement in mobile ad-hoc networks," in *PERCOM '08*. IEEE Computer Society, 2008, pp. 548–553.
[16] C. Frank and K. Römer, "Distributed facility location algorithms for flexible configuration of wireless sensor networks." ser. Lecture Notes in Computer Science, vol. 4549. Springer, 2007, pp. 124–141.
[17] K. Oikonomou, I. Stavrakakis, and A. Xydias, "Scalable service migration in general topologies," in *WOWMOM*, 2008, pp. 1–6.
[18] G. Smaragdakis, N. Laoutaris, K. Oikonomou, I. Stavrakakis, and A. Bestavros, "Distributed Server Migration for Scalable Internet Service Deployment," *IEEE/ACM Transactions on Networking*, (To Appear) 2010.
[19] M. Evans, A. Phippen, G. Mueller, S. Furnell, P. Sanders, and P. Reynolds, "Strategies for content migration on the world wide web," *Internet Research*, vol. 9, no. 1, pp. 25–34, 1999.
[20] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
[21] ——, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239.
[22] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. The MIT Press, September 2001.
[23] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.
[24] R. Solis-Oba, "Approximation algorithms for the k-median problem," in *Efficient Approximation and Online Algorithms*, ser. Lecture Notes in Computer Science, vol. 3484. Springer, 2006, pp. 292–320.