Socially-aware Gateway-based Content Sharing and Backup

Jin Jiang
Dipartimento di Elettronica
Politecnico di Torino, Italy
jin.jiang@polito.it

Claudio Casetti
Dipartimento di Elettronica
Politecnico di Torino, Italy
claudio.casetti@polito.it

ABSTRACT

The amount of data that home users generate, store, and share with their friends via a multitude of devices has grown significantly in the past few years. In our paper, we assume that every household is equipped with a home gateway that stores and manages the data collected by the home users. To accelerate the content sharing and backup for such users, we propose an efficient backup scheme that hinges upon gateway interactions exploiting the users' social networking information. We formulate this problem as a Budgeted Maximum Coverage (BMC) problem and we numerically compute the optimal content backup solution under a synthetic social network scenario. Then, we compare it with two different content placement strategies for gateways with various quota sizes, in a realistic synthetic social network.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*Distributed applications*; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Design, Algorithms, Performance

Keywords

Socially-aware, Content Backup, Home Gateway

1. INTRODUCTION

Home users have become producers, importers, and exporters of large amounts of digital content via a multitude of devices that are all managed independently. As a consequence, there is a clear need for an intelligent content distribution system that can provide a unified content storage and access from within the home and via the Internet, and help the home users exchange data between devices, share it with other users, and assure safe backup.

HomeNets'11, August 15, 2011, Toronto, Ontario, Canada.

In order to realize this vision, a gateway-centric network architecture is proposed, whereby each home is equipped with a gateway and a large number of interconnected devices within the household. The gateway allows any content to be downloaded from outside the household, stored on it, and accessed by satellite devices in communications range of the home gateway. We also assume that, in keeping with the "federated homes" vision, multiple neighboring or remote home gateways can be connected in a collaborative fashion, and can exchange various information.

The intuition that is pursued in our paper is the following: if a gateway is allowed to collect its users' social networking data, such information could be exploited to combine content sharing and content backup. Social data could include (but not be limited to) social contacts and social interests, friends' locations and whether they are in the federated home network or not. Federated gateways could then be designed so as to reserve a part of their storage quota to store content from other gateways belonging to friends from their users' social networks. The content could reflect common interests among such friends. For example, instead of (or in addition to) an anonymous "cloud" backup, a gateway could autonomously choose to upload a set of Mozart concertos to a music-lover friend's gateway. Beside creating a backup of the music files in a trusted location, these could also be enjoyed by the friend, who can access them on his/her own gateway.

Several practical considerations, however, force us to draw a more complicated picture. Firstly, there are gateway selections issues. Choosing a friend's gateway to back up data only because interests match is not a sound policy from a networking point of view. The remote gateway could have poor connectivity or it could be overloaded. The gateway could be located in a far away country (even though friends in social networks are more likely to be in nearby areas [1]). The remote gateway should implement a solid quota management to avoid being swamped by friends' contents.

Additionally, there are architectural details to address: the user must rely on the backed-up content to be kept on the remote gateway (or, at least, it should be notified when the content is about to be deleted). The content should be readily available. A reliable updating policy should also be devised. Lastly, copyright and owndership issues should be regulated.

The architectural issues outlined above are beyond the scope of this work. We chose instead to focus on gateway selection and, specifically, to devise an efficient content placement scheme to determine where to back up the con-

tent from a user's gateway to remote gateways belonging to his/her social friends. As remarked above, placing content replicas "outside" the home (i) consumes transmission bandwidth for uploading the content and (ii) incurs a storage cost on the remote friends' home gateways. So we aim at a strategy that maximizes the friends' utility by trying to match content type and friends' interests while taking into account the bandwidth contraints between gateways as well as the storage space at the remote gateway.

We model this optimization problem as a Budgeted Maximum Coverage (BMC) problem to obtain the optimal content placement solutions under a synthetic social networking scenario, whose set up is also discussed in the paper.

The rest of this work is organized as follows. Section 2 describes the assumptions of our model and the modeling procedure of the content placement problem, which is formulated as an BMC optimization problem. Section 3 introduces the test scenario and presents a procedure to construct a synthetic social network. Also the results derived through simulations using the Gurobi optimizer are presented in Section 3. Finally, Sections 4 and 5, respectively, review some related works and draw some concluding remarks and outline future work.

2. MODEL DESCRIPTION

Our optimization problem aims at maximizing the utilities/benefits of users hosting their friends' backups. To this optimization problem a series of constraints are added, including the boundedness of the total resource capacity of every gateways.

2.1 Assumptions

Consider there are N households in the network, each equipped with one home gateway, hence N is the number of home gateways in the network. A home gateway GW_i $(i = 1, 2, \ldots, N)$ acts as a repository storing content for all users in the corresponding household. It capacity is split into a local data storage (i.e., for data primarily stored by its local users) and into a "friend quota", Q_i, which we define as the available storage capacity for the data uploaded by friends of its users. We define the upload and download bandwidth of the gateway GW_i as $C_i^{(u)}$ and $C_i^{(d)}$, respectively. The bandwidth C_{ih} from gateway GW_i to gateway GW_h is assumed to be:

$$C_{ih} = \frac{min\{C_i^{(u)}, C_h^{(d)}\}}{\alpha(d_{ih})} \tag{1}$$

where $\alpha(d_{ih})$ is a factor depending on the distance d_{ih} between gateway GW_i and GW_h, and $1 \leqslant \alpha(d_{ih}) \leqslant 10$. In Section 3 we will provide a possible definition of $\alpha(d_{ih})$.

We then assume that there are M users in the network. Each user registers himself/herself on the corresponding home gateway, where it accesses/stores the content. For the purpose of identifying which users are registered to which gateway, we define an $N \times M$ matrix $\mathbf{P}$ whose generic element is given by:

$$P_{ij} = \begin{cases} 1 & \text{if } U_j \text{ registered on } GW_i \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where i indicates the gateways, $i = 1, 2, \ldots, N$, while j is the user index, $j = 1, 2, \ldots, M$.

As explained earlier, we assume that a crucial gateway functionality is the capability to collect the social information of its users by extrapolating such data from the social networks they belong to. In particular, we are interested in collecting *user's friend lists* and *user's interests*.

To represent the first dataset, we model the friendship between user U_j and user U_f through a friendship function $F(j, f)$:

$$F(j,f) = \begin{cases} 1 & \text{if } U_j \text{ and } U_f \text{ are friends, } j \neq f; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The friend list E_j of user U_j can thus be denoted as follows:

$$E_j = \{U_f : F(j,f) = 1\} \tag{4}$$

Secondly, the user's interests are mapped from one or more social interest communities that the user belongs to. The degree of user involvement with each such community - which will come to represent the distribution of content type preferences of the user - is captured by the user's *interest vector*, defined as follows. Let I_{jl} denote the interest factor of user U_j in interest type l, with $0 \leqslant I_{jl} \leqslant 1$, $l = 1, 2, \ldots, L$ (L is the size of the interest area, i.e., the total number of interest types considered in our system). Let the interest vector of user U_j be the collection of all interest factors of the user associated with all the interest types, denoted by

$$\overline{I}_j = (I_{j1}, I_{j2}, \ldots, I_{jl}, \ldots, I_{jL}) \tag{5}$$

where $\sum_{l=1}^{L} I_{jl} \triangleq 1 - r^j$. r^j is the probability of the user U_j to be interested in the interest type out of the interest area L. Without loss of generality and in order not to burden the presentation of this problem, we will just assume users to only have interest in the interest types considered in the system, thus, $\sum_{l=1}^{L} I_{jl} = 1$ or $r^j = 0$.

We assume that the content items in the network are finite, i.e., each user may own any number of items out of K possible items. A generic item k, $k = 1, 2, \ldots, K$ has size $D^{(k)}$ and belongs to interest type l. The association between an item and its interest type is assigned according to a uniformly random distribution. For the sake of notation simplicity, we also assume that every user has the same average number of items to share or backup.

2.2 Mapping onto a BMC problem

Our objective is to find a selection of friends from the user's friend list where to back up user's items; such selection should maximize the benefit of the hosting users, i.e., by closely matching his/her interests; and it should maximize the data transfer effectiveness, i.e., by maximizing the bandwidth between the respective gateways.

As previously observed, we cast the optimization problem as a BMC problem. In BMC problems, a collection of sets $S = \{S_1, S_2, \ldots, S_m\}$ with associated costs $\{c_i\}_{i=1}^m$ is defined over a domain of elements $X = \{x_1, x_2, \ldots, x_n\}$ with associated weights $\{w_j\}_{j=1}^n$. The goal is to find a collection of sets $S' \subseteq S$, such that the total cost of elements in S' does not exceed a given budget L, and the total weight of elements covered by S' is maximized. The BMC problem is NP-hard, and [2] presents a $(1 - 1/e)$-approximation algorithm for it.

We assume that gateway GW_i has already collected the user U_j's friend list E_j and the friends' registration information (which friend of U_j is registered on which gateway).

In our case, bins and elements of the BMC problem can be mapped in the following way.

- The *bin* set B_j for user U_j is defined as follows: $B_j = \{b_{j1}, b_{j2}, \ldots, b_{jh}, \ldots, b_{jN}\}$, where bin b_{jh} denotes the set of friends of user U_j who are registered on gateway GW_h, $h = 1, 2, \ldots, N$:

$$b_{jh} = \{U_f \in E_j : P_{hf} = 1\}. \tag{6}$$

  We recall that $P_{hf} = 1$ means that user U_f is registered on gateway GW_h, and that $U_f \in E_j$ means that user U_f is in the friend list of user U_j, so $b_{jh} \subseteq E_j$. The cost $c^{(k)}_{(b_{jh})}$ of selecting the bin b_{jh} is defined as the cost of uploading the content item k of size $D^{(k)}$ onto the gateway GW_h, which can be defined as:

$$c^{(k)}_{(b_{jh})} = D^{(k)} \tag{7}$$

- The *element* set in our problem obviously is the user U_j's friend list E_j.

  For each element/user $U_f \in E_j$ (user U_f is a friend of user U_j), we can define the weight as the benefit $w^{(k)}_{(U_f)}$ that element U_f can obtain when item k is uploaded onto gateway GW_h where U_f is registered. Such benefit will depend both on the interest that U_f will have in the uploaded content, and on how easily accessible that content will be for U_j (i.e., on the bandwidth between the uploader gateway and the hosting gateway). We can thus define $w^{(k)}_{(U_f)}$ as:

$$w^{(k)}_{(U_f)} = I_{fl} \cdot C_{hi} \tag{8}$$

  where I_{fl} denotes the interest of U_f in items of type l which content k belongs to and C_{ih} is the bandwidth from the remote gateway GW_h to the uploader gateway GW_i, on which users U_j and U_f are registered, respectively ($P_{ij} = P_{hf} = 1$).

Our constraint is the gateway friend quota, Q_i, which we recall is the available storage capacity for data uploaded by friends.

Finally, our problem can be formulated as follows:

$$\begin{aligned} &maximize && \textstyle\sum_{k=1}^{K} \sum_{U_f \in E_j} w^{(k)}_{(U_f)} \cdot y^{(k)}_f \\ &subject\ to && \textstyle\sum_{k=1}^{K} c^{(k)}_{(b_{jh})} \cdot x^{(k)}_h \leqslant Q_h \\ & && \textstyle\sum_{P_{hf}=1} x^{(k)}_h \geqslant y^{(k)}_f \\ & && x^{(k)}_h, y^{(k)}_f \in \{0, 1\} \end{aligned} \tag{9}$$

where $x^{(k)}_h = 1$ indicates that gateway GW_h is selected to host a backup of content item k, while $y^{(k)}_f = 1$ means that user U_f is chosen as back-up for the content item k.

2.3 Problem size

The number of Boolean decision variables ($x^{(k)}_h$ and $y^{(k)}_f$) is $O(K\langle N\rangle)$, where $\langle N\rangle$ denotes the average number of the friends per user. And the number of constraints is $O(K\langle N\rangle + M)$. We solve it through the Gurobi solver. The solution time for an instance with approximately $1,000$ gateways, $3,000$ users and an average of 5 content items for each user to share or backup, is about 30 minutes using a 4-core 2.3 GHz system and a 4 GB RAM.

3. BENCHMARKING THE MODEL

In the following, we test the validity of our approach by numerically solving it and deriving the maximal benefit according to eq. (9). The results will be benchmarked against two other, simpler content placement strategies.

Our first problem, however, is the definition of a suitable scenario that must exhibit realistic features of a social network, namely the distribution of friends and their position/distance on the network topology.

3.1 A synthetic social network

To make our simulation scenario more realistic, we need to set up a synthetic social network that shares the basic common properties of real social networks. We choose Facebook as a target, since it is one of most popular and largest online social networking sites nowadays. In particular, we use the findings in [1] and [3] to characterize the network properties and to establish a relationship between geographical distance and friendship probability that matched the one that can be measured in Facebook. We then proceed according to the following three phases.

Phase 1: location and bandwidth assignment.

At first we assign the geographical location information for each home gateway in the scenario. We uniformly distribute the N gateways in an area of $1,000 \times 1,000$ square miles. Next, we compute the geographical distance between any two gateways and we evaluate the bandwidth between them through equation (1). We define the factor $\alpha(d_{ih})$ so that, intuitively, it is small (hence the bandwidth is large) for nearby gateways (up to 0.1 miles apart). Then, we let it grow linearly (hence the bandwidth decreases) up to 1,000 miles, after which we keep it constant. The choice of the values is clearly arbitrary, but it will serve our purpose of introducing a distance-dependent inter-gateway bandwidth. The $\alpha(d_{ih})$ factor is defined as:

$$\alpha(d_{ih}) = \begin{cases} 1 & 0 < d_{ih} \leqslant 0.1 \\ \frac{d_{ih} - 0.1}{111.1} + 1 & 0.1 \leqslant d_{ih} < 1000 \\ 10 & d_{ih} \geqslant 1000 \end{cases} \tag{10}$$

where d_{ih} denotes the beeline distance between gateways GW_i and GW_h.

Phase 2: user assignment.

The next step consists in distributing the M users onto the gateways, in a uniformly random fashion, so that the average number of users per gateway is the same. Each user will therefore have a geographical location according to the home gateway on which it has been registered. We then compute the geographical distances between each pair of users.

Phase 3: friendship grooming.

The final step is the crucial one. The user graph we have constructed in the previous two steps has some degree of plausibility but it does not reflect yet any properties of actual social networks. In particular, if we established friendship between users picked at random on the graph, we would lose the typical locality that is exhibited by social networks, where most online friends tend to live in nearby areas, due to habits, employment or existing relationships. In order

to create a plausible synthetic social graph based on the geographical location we were inspired by [4]. Such work presents a stochastic model for spatially embedded social networks based on the ideas of spatial interaction models. In it, each user is assigned an identical budget, and if two users establish a friendship link, the two users should both consume a cost amount from their respective budgets. When a user's budget reaches zero, the user will not be assigned any more friends. We call this cost a "friendship cost" and following [4] we define it depending on the geographical distance as follows:

$$c_{F(j,f)=1} = \gamma ln(d_{jf}) + const \tag{11}$$

where d_{jf} denotes the geographical distance between the user U_j and U_f, and $F(j,f) = 1$ means user U_j and U_f are friends, as defined earlier. In our study, we choose $\gamma = 1.05$ and $const = 1$ following the suggested values in [4].

Two details are still missing though: if every user has the same budget, the friendship graph will not resemble a social graph. And we still need to factor in the information on the distance distribution among friends in a social network.

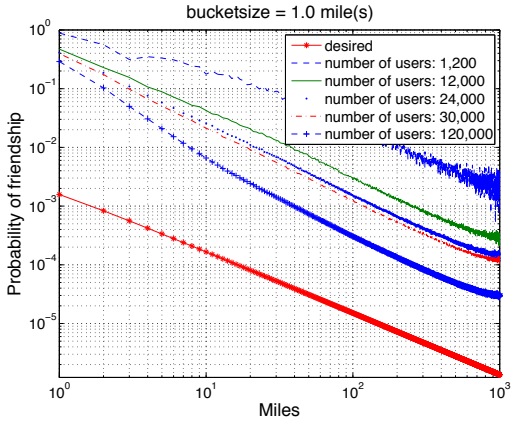

Figure 1: Probability of friendship as a function of distance.

To address the first concern, we change the budget assignment so that users initially receive a randomly assigned budget which follows a power-law distribution with exponent -1.5. Thus, we can let the degree distribution in the social graph (i.e., the number of friends per user) follow a power-law distribution as in realistic social networks. Also we can adjust the average budget value to obtain the desired average node degree.

Secondly, to tie in the distance distribution among friends we follow the findings in [1], where the probability of a friendship link between two users in Facebook is empirically determined given the two users' geographical distance, as follows:

$$p_{F(j,f)=1} = 0.0019 \times (d_{jf} + 0.196)^{-1.05} \tag{12}$$

where $p_{F(j,f)=1}$ denotes the probability that user U_j and U_f are friends.

The resulting social graph is then costructed through the following steps: (i) randomly pick a pair of users among those defined in Phase 2; (ii) use the friendship probability in (12) to determine whether to establish a friendship link among them; (iii) if the link is established, the budgets of U_j and U_f are decreased by the amount in (11).

The above procedure is repeated until all budgets of all users have been consumed[1].

In order to validate this approach, we compare the resulting friendship distribution as a function of user distance with the empirical results in [1]. The comparison is shown in Figure 1. The curve labeled "desired" is plotted following the results in [1], where about one million Facebook users were sampled. The other curves show the distributions derived from our procedure for varying number of users. Although we were yet unable to find an optimized implementation of our procedure that allows us to handle millions of users, the trend shown in Figure 1 is a clear indication that the distribution we obtain can be considered plausible.

Table 1: Parameters used in the scenario

Parameter Name	Notation	Value
Number of Gateways	N	1,000
Number of Users	M	1,200 - 3,000
Number of content items	K	6,000 - 15,000
Items to backup per user		5
Number of interest types	L	10
Content item size	$D^{(k)}$	$10MB$
Uplink bandwidth	$C_i^{(u)}$	$4Mb/s$
Downlink bandwidth	$C_i^{(d)}$	$8Mb/s$

3.2 Simulation scenario and results

We consider a realistic synthetic social network assembled following the procedure outlined in the previous subsection, and with the parameters listed in Table 1. As previously described, we assume that each user has an interest vector on the different content types, and the user's interest vector is shared with his/her friends in the whole network.

Then we consider three different content placement strategies and compare the performance of these strategies in the same scenario. The first strategy is the *joint* optimization method, described in Section 2, in which the friends are selected who have the largest interest in the corresponding content item and are connected with the highest bandwidth (assuming their quota is not used up). The second strategy is the *bandwidth-based* method, in which friends reachable through gateways with the highest bandwidth will be selected, regardless of their interest in the uploaded items. The last one is the *random* method, in which the user just randomly chooses up to 10 friends to share the content item with, as long as the friends have enough quota to store the item, not considering any other factors.

We find the optimal allocation for the first two strategies through the Gurobi solver [5], which uses a variant of the branch-and-cut algorithm. Solving the budgeted maximum coverage problem in eq. (9) yields the optimal joint content item placement, i.e., the set of candidate home gateways to select for each content item, as well as the optimal benefit value that each user can obtain by being selected. The optimal bandwidth-based placement is obtained again from eq. (9) by changing the benefit definition into $w_{(U_f)}^{(k)} = C_{ih}$. Using the obtained optimal total benefit, or weight, we com-

[1]up to a minimum tolerance value, since it is unlikely that a budget becomes exactly zero

pute the average per-user benefit obtained in the system, for various quota constraints, shown in Figures 2 and 3, for $M = 1200$ and $M = 3000$ users, respectively (resulting in 1.2 and 3 users per gateway). Even though the average benefit is lower when 3000 users are considered (due to the larger number of users per gateway sharing the same quota), the advantage of finding an optimal allocation for backed-up content depending on interest and bandwidth is clearly visible in both plots. Such advantage amounts to twice the benefit obtained with a bandwidth-based strategy only, and to ten times the benefit of a random back-up selection.

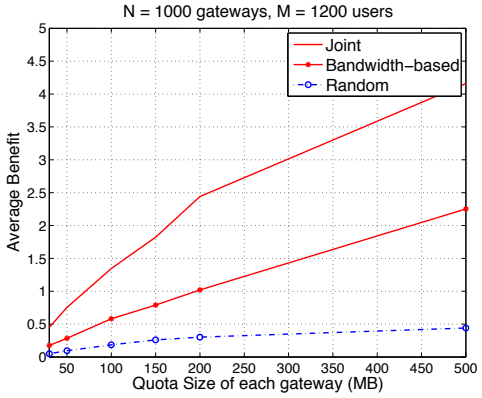

Figure 2: Average benefit per user as a function of gateway quotas, $M = 1200$

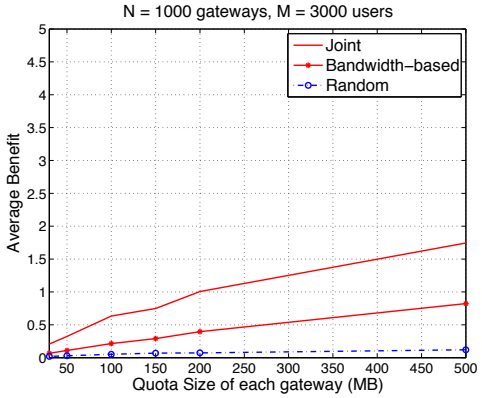

Figure 3: Average benefit per user as a function of gateway quotas, $M = 3000$

4. RELATED WORK

Several research efforts have considered the problem of managing data across replicas. Solutions such as PRACTI [6] and Cimbiosys [7] additionally provide partial replication capabilities to better utilize the available storage capabilities. Podbase [8] provides a framework for automatically ensuring that multiple copies are stored across devices. But these works do not consider or exploit the effects of social networking.

Also there are some works on content placement exploiting information from social networking. The work in [9] proposes ContentPlace, which is a social-oriented framework for data dissemination. ContentPlace 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 [10] where it is shown that mobility and cooperative content replication strategies can help bridge social groups. Another relevant work on an efficient social-aware content placement in opportunistic networks is [11] in which the authors model the content placement as the facility location problem.

5. CONCLUSIONS

The goal of this work is to look beyond traditional approaches for content sharing and backup involving home networks. In particular, we believe the user's utility accounting for user's interests and network bandwidth can be maximized by placing the content "outside the home" in a cloud formed by other home networks and exploiting the user's social networking information. We formulated this optimization problem as a budgeted maximum coverage problem and solved it numerically in a synthetic social network. Finally we evaluated and compared the performance of three different content placement strategies across different home gateway quota cases and showed that the joint interest/bandwidth optimization strategy is superior to other one.

Our ongoing work prominently features the search for heuristic algorithms that approximate the joint optimization strategy taking into account the dynamic nature of users' social networking information.

6. ACKNOWLEDGMENTS

This work was partly funded by the European Union, through its 7th Framework Programme for Research (FP7), under grant agreement 258378 - FIGARO project.

7. REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. *19th ACM WWW*, April 2010.

[2] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[3] C. T. B. Minas Gjoka, Maciej Kurant and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. *IEEE INFOCOM 2010*, March 2010.

[4] J. Illenberger, G. Flötteröd, M. Kowald, and K. Nagel. A model for spatially embedded social networks. *12th International Conference on Travel Behaviour Research*, December 2009.

[5] Gurobi optimizer. `http://www.gurobi.com/html/products.html`.

[6] N. Belaramani, M. Dahlin, L. Gao, A. Nayate, A. Venkataramani, P. Yalagandula, and J. Zheng. PRACTI replication. *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, May 2006.

[7] V. Ramasubramanian, T. Rodeheffer, D. Terry, M. Walraed-Sullivan, T. Wobber, C. Marshall, and A. Vahdat. Cimbiosys: A platform for content-based partial replication. *USENIX Symposium on Networked Systems Design and Implementation*, August 2009.

[8] A. Post, P. Kuznetsov, and P. Druschel. PodBase: Transparent storage management for personal devices. *USENIX International Workshop on Peer-to-peer Systems*, February 2008.

[9] C. Boldrini, M. Conti, and A. Passarella. ContentPlace: social-aware data dissemination in opportunistic networks. *ACM MSWiM 2008*, ACM, October 2008.

[10] E. Jaho and I. Stavrakakis. Joint interest-and locality-aware content dissemination in social networks. *IEEE/IFIP WONS 2009*, February 2009.

[11] P. Pantazopoulos, I. Stavrakakis, A. Passarella, and M. Conti. Efficient Social-aware Content Placement in Opportunistic Networks. *IEEE/IFIP WONS 2010*, February 2010.