Challenges in Personalizing and Decentralizing the Web: An Overview of GOSSPLE*

Anne-Marie Kermarrec

INRIA, Rennes Bretagne-Atlantique, France
Anne-Marie.Kermarrec@inria.fr

Abstract. Social networks and collaborative tagging systems have taken off at an unexpected scale and speed (Facebook, YouTube, Flickr, Last.fm, Delicious, etc). Web content is now generated by you, me, our friends and millions of others. This represents a revolution in usage and a great opportunity to leverage collaborative knowledge to enhance the user's Internet experience. The GOSSPLE project aims at precisely achieving this: automatically capturing affinities between users that are potentially unknown yet share similar interests, or exhibiting similar behaviors on the Web. This fully personalizes the search process, increasing the ability of a user to find relevant content. This personalization calls for decentralization. (1) Centralized servers might dissuade users from generating new content for they expose their privacy and represent a single point of attack. (2) The amount of information to store grows exponentially with the size of the system and centralized systems cannot sustain storing a growing amount of data at a user granularity. We believe that the salvation can only come from a fully decentralized user centric approach where every participant is entrusted to harvest the Web with information relevant to her own activity. This poses a number of scientific challenges: How to discover similar users, how to define the relevant metrics for such personalization, how to preserve privacy when needed, how to deal with free-riders and misheavior and how to manage efficiently a growing amount of data.

1 Introduction

While the Internet has fully moved into homes, creating tremendous opportunities to exploit the huge amount of resources at the edge of the network, the Web has changed dramatically over the past years. There has been an exponential growth of user-generated content (Flickr, Youtube, Delicious, ...) and a spectacular development of social networks (Twitter, FaceBook, etc). This represents a fantastic potential in leveraging such kinds of information about the users: their circles of friends, their interests, their activities, the content they generate. This also reveals striking evidence that navigating the Internet goes beyond traditional search engines. New and powerful tools that could empower individuals in ways that the Internet search will never be able do are required.

The objective of GOSSPLE is to provide an innovative and fully decentralized approach to navigating the digital information universe by placing *users affinities and preferences* at the heart of the search process. Where traditional search engines fail to provide information unless it is properly indexed, GOSSPLE will seek the information where it ultimately is: *at the user*.

* This work is supported by the ERC Starting Grant GOSSPLE number 204742.

R. Guerraoui and F. Petit (Eds.): SSS 2009, LNCS 5873, pp. 1–16, 2009.

GOSSPLE aims at capturing the interactions and affinities on the fly and fully leveraging the huge resource potential available on edge nodes, to efficiently search, dynamically index and asynchronously disseminate and recommend information to interested users based on their preferences. Building on the peer to peer communication paradigm and harnessing the power of gossip-based algorithms, GOSSPLE aims at personalizing Web navigation, by means of a fully decentralized solution, for the sake of scalability and privacy.

A number of technical challenges underlie GOSSPLE and its objective of combining personalization and decentralization:

- **Personalization:** GOSSPLE should address appropriate metrics to compute distances between users and identify and capture the affinities between users.
- **Scalability:** GOSSPLE should provide scalable mechanisms to deal with a huge and growing amount of information.
- **Privacy:** while entrusting users to hold and maintain their personal data give them full control on them, further mechanisms are required in GOSSPLE to leverage personal information and detect affinities between user without exposing personal information about the requests of a user or the content she generates.
- **Support for misbehavior:** while fully decentralized approaches buy scalability, they remove any form of central authority, leaving holes for misbehavior: GOSSPLE should tackle the whole range of misbehavior from attempts to free-ride the system, to attempts to try to exploit it (through spamming for example) and even hurt it with Byzantine behaviors.

The rest of the paper provides the context and motivation (Section 2), the technical challenges (Section 3), the scientific background (Section 4) before concluding and providing the current status of the GOSSPLE project.

2 Time for a Navigation Shift in the Internet

The past decade has witnessed a dramatic scale shift in the area of distributed computing. Meanwhile, the Internet has entered our homes together with various kinds of digital assets. This has resulted into a radical change in the way people are communicating, companies are organized and data is managed all over the world. Social networking in the forms of social networks (Facebook, Twitter) or folksonomies (Delicious, Flickr) has taken off at an unexpected scale. The Internet we are now looking at is composed of millions of computing devices and as many users, generating contents at a high speed, Terabytes of dynamic data, scattered all over the world, shared, disseminated and searched for.

2.1 Personalized Navigation within the Internet

Although computer science in general and more specifically distributed computing has gradually taken into account this digital revolution, we now have reached the point where incremental changes are no longer sustainable. Traditional search engines are performing extremely well but do hardly encompass alternative and very dynamic sources

of information such as user-generated contents, blogs, peer-to-peer file-sharing systems instant messaging as well as content distribution frameworks. This is mainly due to their lack of adaptivity to dynamics and their not taking into account correlations between contents and users preferences. They are also limited by their reliance upon centralized indexing: they periodically scan the whole web, build an index in their data centre, then distribute it back out to smaller centres that respond to queries. Typically, corporate pages are visited frequently while individual information may be visited rarely: *the individual is at a disadvantage*. This reveals striking evidence that complementary and novel fully decentralized alternatives to traditional search engines are now required to capture the dynamic, collaborative and heterogeneous nature of the digital universe as well as to to leverage individual preferences and social affinities.

2.2 Illustration: Looking for a Baby-Sitter

To illustrate the inadequacy of state of the art solutions, let us consider the following concrete example. Following a long stay in the UK, a French family is looking for an English speaking student who would be willing to trade baby-sitting hours against accommodation, say in the city of Rennes to allow kids to keep up with English. Given the high number of students in Rennes, there is no doubt that such an offer would be of interest for many English speaking students.

Yet, satisfying this simple, slightly unusual, request is challenging and in fact almost impossible. The most natural way for the family to find a match is to launch a Google request "Baby-sitter anglophone Rennes"[1]. The first hits on Google lead to baby-sitting services, student announces, including different geographical areas and has nothing to do with English speaking. All subsequent reformulated requests, in French or English, lead to equally unsatisfactory results. Yet, would this family be able to reach all English speaking students in Rennes, there will definitely be some candidates.

The data is clearly out there but it is difficult to achieve the match between the offer and the supply. If the offer effectively exists in some proper indexed form, even though a search engine forces to continuously probe the system, it will probably achieve the match eventually. Alternative sites such as Craigslist, a centralized network of online communities featuring free classified advertisements, extremely popular in the US, could also be used in this case, provided that the user follows the imposed structure. However, if the offer does not exist in the proper indexed form, current technology simply does not fit. This is mostly due to the fact that *baby-sitter* is mainly associated with *daycare* or local baby sitting companies. None of the family Facebook buddies can help either as known of them has ever looked for an English-speaking baby-sitter. The best solution would be for the family to post a request on some mailing list or appropriate forum gathering the potential candidate baby-sitters and wait for the responses.

Now, consider Alice living in Strasbourg, who has looked for a similar deal for her kids. Alice is lucky enough to discover through a (real-life) friend that primary school teaching assistants are a very good match for they have the same working hours as kids and tend to enjoy living with a family. If Alice associates *baby-sitter* with *teaching assistant* in the system and if the French family above is able to leverage this information,

[1] "English speaking baby-sitter Rennes".

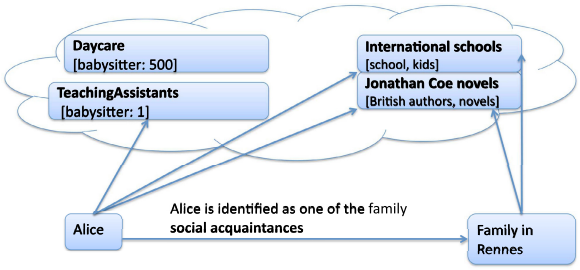

Fig. 1. Babysitter example: while the association between babysitter and daycare dominates, Alice associates babysitter to teaching assistant. The goal of GOSSPLE is to establish a connection between Alice and the French family in Rennes so that it could benefit from Alice's association.

the request can be successful. The goal of GOSSPLE is to establish such a connection, called an implicit social link, between Alice and that French family in Rennes. Note they do not need to know each other. Yet their past history of French people leaving in an English speaking country, their interest in English novels and International school for example, could be conveyed in their online behavior and automatically captured by a system. This is illustrated on Figure 1.

2.3 Where GOSSPLE Comes into Place

In fact, the collaborative and social nature of the Internet is leveraged in many social systems [28] such as delicious, Twitter, Facebook, Twine or Orkut to cite a few. Such systems connect users sharing interests, professional or social, and enable them to share data, blogs, etc. Their functioning is however hurt by the dynamic nature of users behavior. Some users get connected, loose interest and remain connected without participating. Also, the user feedback is hardly leveraged and while the blog feature is widely used, search is mostly absent. Similarly, the semantic Web improves automation through machine understandable descriptions [11]. Yet, such tools mostly rely on static structures. Above all, all those systems remain centralized. This an issue for two main reasons: scalability and privacy. An efficient personalization mechanism requires to store a large amount of data per user and maintain it, potentially limiting the scalability of the system and hurting the desire of users to preserve their personal information. In addition, centralized systems are more vulnerable to denial of service attacks such as the one observed in August 2009 on Twitter, Facebook and LiveJournal.

To cope with dynamics and the huge amount of information that need to be managed on a per user basis, entrusting each user with discovering and managing the data relevant to her is the solution to both scalability and privacy preservation.

GOSSPLE stems from the observation that social connections can be leveraged by a system to collaboratively help Web search and recommendation. Yet such social connections need not to be explicitly established as in social networks ala Facebook. Instead the system should capture such social connections and discover relevant users. As opposed to globally harvest and organize the Web, the basic idea behind GOSSPLE is that each user is in charge of harvesting the network in her own personalized way.

Coming back to our example, even if the answer to the request actually does not exist as such (say no foreign student has figured out that some families would offer such a deal), GOSSPLE would actually enable to dynamically *attract it*. There are several ways this could be achieved by GOSSPLE, by expanding the query in a relevant manner or by having the request navigate in the network to the right places. With GOSSPLE, the family would gradually get connected to relevant matching users typically representing adequate communities (say English speaking people in Rennes). Then the object would dynamically turn into an ad, in a sense *creating the need* and subsequently the matches. In turn, potential response objects would travel back to the family acquaintances in the form of notifications or ads, and subsequently create the need for other related families (those who wouldn't have thought of the deal but actually like the idea). At the heart of this procedure lies dynamic overlays based on *users affinities and preferences*. This goes far beyond discovering indexed data. All along, the connection procedures, both sides, will be continuously fed by the feedback from the users to refine the quality of the connections, as well as by recommendations on possibly matching objects from other users with similar preferences on similar requests. The interacting model is inherently collaborative, asynchronous and iterative.

Obviously, this example is not meant to restrict the usage of GOSSPLE to this application. However, we believe that the simple scenario illustrates the dynamic and collaborative navigation idea. These, implemented in a fully-decentralized manner, can be applied to a large spectrum of applications (content sharing, dissemination, instant messaging, RSS feeds, or virtual communities).

3 The GOSSPLE Challenges

The existing technology of distributed and personalized search is in its infancy. We are reaching the limits of what we could call the "Google style" of problem solving: periodically cull all the pages on the web into their data centre, index them, and then answer queries for pages for some period of time. So far, the information space has mostly been composed of Web pages, indexation ruled by search engines and navigation ensured mostly manually by the users, largely favouring the "mass". Effectively, the page rank algorithms of Google-like systems favour popular pages. Although GOSSPLE does not come as a replacement of such engines but rather as a complementary tool, it provides a fresh look at the information space management and favour communities at a disadvantage. More specifically it offers a new way to navigate the digital space.

The GOSSPLE's challenge is to provide the following features in a fully decentralized way.

1. Full-fledged personalization
2. Scalable management of the information space
3. Privacy-aware implementation
4. Resilience to misbehavior

3.1 A Network of Affinities

We are seeking search solutions leveraging the live nature of the data and the collaborative nature of its users. GOSSPLE exploits the social dimension of the Internet to get

"related" users indirectly connected and refine each other's filtering procedures through implicit preferences. The network will be organized around such preferences and affinities between users. This will provide a radically different approach to managing digital assets, navigating within the Internet and bringing new dimensions for collaborative applications. Such a network of affinities is at the heart of GOSSPLE and represents the first challenge of Gossple. Providing each user with a personalized view of the network requires solving several issues:

- **Sampling the network:** the second challenge is to be able to discover such users. This is particularly challenging in a fully decentralized system where no entity has a global knowledge of the system and is able to make a match between similar users. A related issue is to connect all GOSSPLE users in a connected mesh: although a user should be connected to similar users, she should be able to navigate the whole network if needed.
- **Affinity metrics:** the first challenge is to be able to identify the fact that two users share similar interests. This requires to compute a distance between users and can depend on the content they generate, their past activity, the feedback they provide, the application they are running, etc.
- **Coping with dynamics:** the third challenge is to be able to maintain a personalized network up-to-date and to take into account the changes and the dynamics of the system with respect to the users, the data, the changes in the interests or the activity of the users.

GOSSPLE will heavily rely on peer to peer overlay networks to achieve personalization of the network. Basically, GOSSPLE will manage a large set of GOSSPLE peers (users, items, etc). More specifically, we envision a basic layer where all potential nodes are, at least temporarily, connected and maintained despite dynamics in content and connectivity patterns, providing gateways and efficient routing to higher level overlays (See Figure 2). At the basic abstraction layer, a GOSSPLE peer represents a machine connected to the Internet. The same physical computer may host several logical GOSSPLE peers: the request of the family, a user in a virtual community, a file, etc. A major GOSSPLE challenge is to build, on top of its basic layer, many overlay networks that will dynamically evolve, based on users affinities and common preferences.

GOSSPLE will leverage the sampling features of gossip-based protocols to provide users with the ability to sample the network and identify similar users.

Figure 2 conveys an example of a federation of overlays as we foresee it in GOSSPLE. The bottom layer ensures connectivity, on top of which the federation of overlays is maintained. Each GOSSPLE peer associated with a user may be part of one or several sub-overlay networks, whose nature may vary depending on the functionalities required by the application they are running. This amounts to having a physical peer running many instances of different P2P overlay networks. Yet, a fair amount of information may be shared between these instances. We will investigate the mutualization of the state associated with each overlay in order to limit the overhead for a similar, or even better, performance. More specifically, we will identify for each overlay the application-dependent connections, which will have to be maintained independently of other sub-networks such as the "closest" peers according to the "affinity" metric.

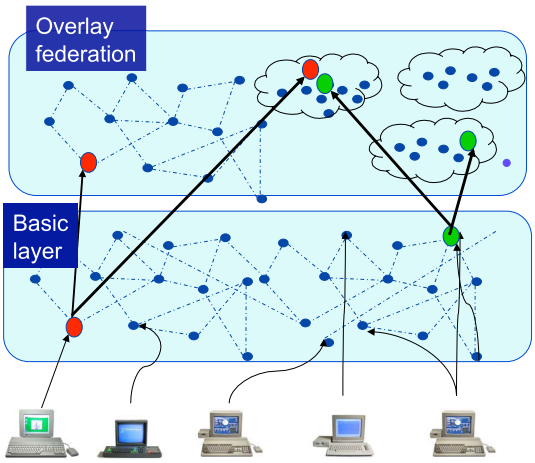

Fig. 2. Gossple Overlay Federation

Identifying the relevant users requires appropriate metrics to compute a distance between users. The refinement of connections between peers is crucial: each GOSSPLE peer will keep in its *personalized view* of the system for a given overlay a set of acquainted peers. In most cases, this choice is done on a peer basis, that is depending on its own characteristics. The correlation between all the peers present in the view could also be exploited to cover as much as possible all the range of interests of a user. All these aspects should be investigated in the project.

3.2 Scalable Data Management

A scalable personalization of the network, operating a navigation shift on the Internet, calls for a fully decentralized system and requires the following features:

- **Efficient management of personal information:** this refers to the amount and the type of personal data that should be stored per user and exchanged between users in order to evaluate the proximity of interests between users and achieve personalization.
- **Efficient search, recommendation and navigation algorithms:** this refers to the algorithms to search content, process queries, implement efficient notification mechanisms, routing features, etc.

Identifying the relevant discovery space, the granularity of the search protocol and data representation are crucial to the design of an efficient digital navigation. The navigation criteria should be simple and flexible enough to preserve the efficiency and simplicity of an underlying gossip-based discovery protocol. A related issue is the trade-off between expressiveness and exhaustivity. Expressiveness refers to the accuracy of a request formulation (exact search, keyword-based search, range queries), or the *quality* of a request. This is highly dependent on the number of dimensions of the search space,

the type of query, the correlation between various attributes of the request. The degree of exhaustivity refers to the accuracy with respect to *quantity*.

A key aspect of GOSSPLE is to capture the commonalities and preferences of users from their matching refinements and then leverage these for efficient navigation. This is crucial to genuinely exploit the collaborative nature of the Internet. GOSSPLE will integrate the feedback from the users in the navigation process through recommendation mechanisms. The acquaintances between related users may take the form of recommendations, as in real life, and the navigation protocol should take those as direct inputs to refine the search either directly or indirectly through specific overlays. These aspects have been so far mostly ignored in the distributed system community and specific mechanisms, simple enough for the user, and not disruptive for the system, need to be investigated.

There are many connections with the information retrieval community. However, most approaches are centralized, complex and require a large amount of knowledge of the whole system. GOSSPLE borrows from this community to represent and track similarities between data and/or users.

3.3 Preserving Privacy

Apart from the fact that centralized systems may be subject to DOS attacks, one of the main motivations to provide a fully decentralized system is to fulfil the need for privacy of users and fight their fear (or the real risk) of the *Big Brother Syndrome*. In the realm of recent developments of social networks, the associated companies have consistently shown their eagerness to exploit personal information: in 2009, Facebook tried to change its terms of use so that any content ever published on Facebook was doomed to a perpetual licence to Facebook. Likewise in 2007, Facebook proposed a feature called beacon to expose Web navigation history of users [2]. Similarly, many personal information are stored by Google [3].

A fully decentralized system avoids such issues as no single entity keeps the control of personal data. Instead, the users are in charge of managing such data themselves. In order to get the most of users communities, personal information must be disclosed to some extent. Yet, the association between a user and her personal information is not always required. The challenge here, with respect to privacy, is to ensure that personal information can be fully leveraged while masking the association between user profile and identity whenever this is required.

GOSSPLE leverages this fact by masking the association between a user and her information whenever this is possible. GOSSPLE will also include a lightweight mechanism to track potential intruders, including colluding ones.

3.4 Fighting Misbehavior

Fully decentralized systems are particularly vulnerable to misbehavior, the very fact that there is no central control authority allow users to misbehave with impunity, ranging

[2] Note that those proposals did not get through due to users reaction.

[3] To further illustrate this, the launching of Google Latitude on the iPhone, a location-based social network, in July 2009, raised many concerns with respect to privacy. Indeed, many people are extremely reluctant to disclose people whereabouts.

from free-riding behaviors, to malicious ones. Fighting such misbehavior is of the utmost importance for a wide adoption of a system.

Several angles can be investigated:

- Measuring the degree of collaboration in order to characterize the benefit of a user with respect to her contribution
- Detecting misbehavior
- Punishing misbehaving nodes thus creating an incentive to non-malicious behaviors

Load balancing, referring to the fact that the load is evenly shared between participating entities has been at the heart of the design of P2P systems to ensure scalability regardless of the capacities of peers. *Fairness* has not. In this context fairness is related to the ratio between the benefit a peer gets from the system from its contribution. We mean by a fair system one in which peers contribute to the system proportionally to the benefit they get. This is crucial for a collaborative system to provide incentives to contribute. The fact that fairness has been ignored so far is mostly due to the low-level nature of distributed systems, where the perception by a user is not prevalent. This is no longer the case because users and machines are closely related, now more than ever. A user does not want a software to store data for others or use her bandwidth without being rewarded to a certain extent for this. Should users perceive that they contribute to the system more than what they get out of it, they could decide to get disconnected. Thus, an unfair distribution of the workload can lead to increasing artificially the system dynamics and impact the reliability and scalability of a decentralized system. This is particularly important in GOSSPLE where inputs from the users and their affinities are prevalent.

Ensuring fairness implies characterizing the load, being able to measure it, and devising adaptive mechanisms to account for it. Fairness also intrinsically limits the impact of selfish (free-riders) users. Yet, some users may exhibit some arbitrary behavior, voluntarily or not. Clearly, GOSSPLE might suffer from the same potential attacks as a traditional P2P system [4]. In addition, the misbehavior might also target the data that are exchanged in the system in order to personalize the system. Indeed, GOSSPLE introduces some specificities in this area related to the targeted applications such as false recommendations, wrong feedback or stale objects.

4 Background: Peer to Peer, Gossip and the Small World Nature of the Internet

Decentralization is a core characteristic of GOSSPLE. In this section, we provide the networking background on which GOSSPLE will heavily rely.

Traditional structured and unstructured overlays exhibit almost orthogonal properties and are complementary with respect to locating data in a large-scale system. Structured overlays associate keys with nodes and provide an exact match interface. This approach is highly efficient when the exact identifier of an item is known but not as straightforward when it comes to performing a *range query* or a *keyword-based search*. In addition, the maintenance cost of a structured overlay may be high in dynamic environments where the peers leave and join the system frequently. On the other hand, unstructured

networks handle range queries and keyword searches more easily and are highly adaptive to dynamic environments. In particular, the inherent randomness of gossip-based protocols makes their corresponding unstructured networks ideal for scalable information dissemination. However, they tend to generate a large number of messages for each search request as they do not recall any history. Besides, they do not always guarantee an exhaustive search.

We aim at taking the best of all worlds, by combining structured and unstructured overlays within GOSSPLE. More specifically, we will make use of a gossip-based protocol for basic navigation, combined with structured networks derived from the affinities of users.

Self-* emerging structures. Current search engines are mostly centralized [4]. Not only do we aim at revolutionizing navigation, but we also believe that it is no longer conceivable to rely on a few companies to index the digital world [5]. The total absence of centralization is the key to both scalability and privacy preservation. A fully decentralized system, as envisioned in GOSSPLE is sustainable if and only if it is able to be self-organizing, self-healing, self-parametrizing and self-managing. To this end, GOSSPLE will harness the power of gossip-based algorithms, strongly rely on the scalable peer to peer communication paradigm and overlay networks.

Connectivity: Peer to peer communication paradigm. In peer to peer (P2P) systems, each node may be both a client and a server and takes individual decisions based on an extremely restricted knowledge of the network. Yet expected global system properties emerge. This makes P2P computing robust, self-organizing and scalable. Following this model, nodes organize in a logical (overlay) network, structured or not, on top of a physical network (typically the Internet). Many such overlays have been proposed in the past five years [37,32]. Yet, real deployments remain limited and their potential goes far beyond file sharing, voice over IP or content distribution. In GOSSPLE, we step away from general-purpose overlay networks and consider dynamic application-tailored collaborative overlays.

Navigability: Small-world networks. Small world networks have been introduced as an analytical way of modelling the *six degrees of separation* [26] stating that two random individuals are separated by short chains of acquaintances that can be discovered. When applied to computing networks, the small world phenomenon [23] is defined as the combination of a high degree of clustering, small diameter in the connection graph and navigability properties. Such a model matches pretty well the real interactions between humans and more specifically between users over the Internet. A small-world network can be defined as a system where each node in a mesh knows its *closest*[6] neighbors and has additional shortcuts in the graph. The asymptotic routing performance depends on the way shortcuts are chosen (random [44] or following a specific distribution (d-harmonic) [23]). Kleinberg [23] determined the magnitude order of this routing complexity results in such networks. This work has been of the up-most importance in the community, leading to a full range of works.

[4] Obviously central servers in this context refer to huge data centres.

[5] One can imagine the impact of Google falling apart.

[6] The proximity metric may be application-dependent.

Dynamicity: Gossip-based networking. Gossip-based protocols implement the P2P communication paradigm in an unstructured manner. Inspired by the spreading of rumours or epidemics, these protocols are very powerful for disseminating information and quickly discovering acquaintances between users. Their implementation typically relies on a periodic peer-wise exchange of information. It turns out that depending on the peer chosen locally for the interaction and the information exchanged, gossip-based protocols can be used to build and maintain arbitrary structures. As such, gossip-based protocols are attractive for developing large-scale distributed systems and do have a substantial power. They combine convergent behavior, ability to let emergent structures appear, simplicity of programming and deployment. They also impose a bounded load on participants, are independent of the underlying topology and are robust to network disruptions and continuous changes. Gossip-based protocols will constitute a basic building block for the design and implementation of GOSSPLE.

In short, a generic version of a gossip-based protocol, consists in having each peer run periodically a protocol that can be fully characterized by the three following parameters [21]: *(i) Peer selection* refers to the peer selected for the gossip exchange. Each peer has an extremely limited knowledge of the system (list of other peers) and selects a peer from this *view* of the system; *(ii) Data exchanged* refers to the nature of the data exchanged during the gossip interaction. This is highly application-dependent; *(iii) Data processing* refers to the computation operated on the data after the exchange. Again, data processing is highly application dependent. This simple algorithm and its associated set of parameters are surprisingly powerful and can be applied in a wide variety of settings. More specifically, when the data exchanged is related to peer themselves, this provides a generic tool to build and maintain large-scale overlay networks, structured, unstructured, random, or clustered [15,19]. They also cope extremely well with network dynamics. For example, more than 70% of the nodes are required to be down for a network to become partitioned [21]. When the data exchanged is related to information to be disseminated, this provides a scalable and reliable dissemination system [15].

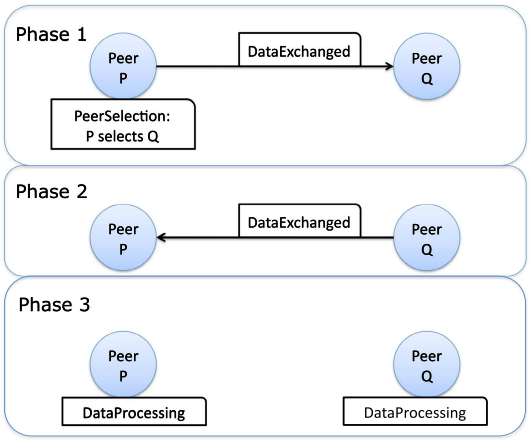

Fig. 3. Phases of a gossip initiated by Peer P: P picks Q among its neighbors (its view)

Distributed computations can also be implemented by simply tuning adequately the data exchanged and data processing parameters [20,25]. Gossip-based protocols have also been used to create clustered overlays optimized with respect to application-specific metrics [42,41]. To illustrate this further, epidemic protocols may be used to construct P2P overlay networks achieving graph properties very close to those of random graphs [21]. The protocol is illustrated on Figure 3.

These protocols scale extremely well and are closely matched to the style of social networking problems GOSSPLE targets. In GOSSPLE, we will go one step further to explore their huge potential over the Internet and in particular consider them with respect to arbitrary metrics.

5 Personalizing the Web: Related Projects

The related work in networks has been presented above. In this section, we provide a brief overview of the work that has been conducted to personalize the Web. Since the Web has been acknowledged as a read-write platform with growing user-generated content, a lot of research has tried to leverage this in many areas [34,24,40]. Yet, to the best or our knowledge no existing work combines personalization, decentralization, privacy and resilience to misbehavior.

Personalized search. The social semantics between users exhibits a huge potential to leveraging social connections should they be explicit through social networks connections or implicit trough similar tagging behavior. One example of system leveraging explicit social connections is Peerspective [27] where the search results of a user's skype buddies are used for the user subsequent search operations. Yet, as pointed out in [9], the utility of the information gathered from such networks turns out however to be very limited. We believe that there is much more to leverage in unknown social acquaintances or user activities such as user's query histories [36], browsing histories [38], and tagging behaviors [31].

In the context of top-k processing, the notion of user affinity has been often discussed [33,3], yet, most personalized approaches are centralized such as [33] or [2]. In the context of query expansion, collecting and exploiting information about the past activity of a user has been considered in [12,22]. The work presented in [8] is a first step to personalization through social relation: the scoring model is personalized, the associated query expansion mechanism is not.

Finally, there have been several user-centric approaches in the area of search, and recommendation [30,47,45,17,9,29,46,39,7,18,49], as well as query expansion [48]. None is decentralized though.

Decentralized approaches. The closest work with respect to distributed systems are semantic overlays, relying on the peer-to-peer communication paradigm. These systems [14,35,6,16] cluster peers hosting similar data or interested in similar topics [43] in order to improve the efficiency of query resolution in peer-to-peer data sharing systems. Their focus is nevertheless mainly on exploiting similarities to locate objects in a distributed data repository. None of these approaches attempt to discover social connections between peers.

Metrics. There has been a lot of work, mostly in the area of information retrieval on personalization metrics to measure the distance between tags or items in collaborative systems, and folksonomies in particular. These include *co-occurrence count* [30], *cosine similarity* to compute distance between users [47,49] or tags [13,46], *edit-distance* [45] and *relative centrality*. Yet, there are still many application-dependent metrics that should be considered.

Finally, recommendation systems ([1] for example) have been proposed and analyzed from a theoretical standpoint, there are yet to be put in practice in a decentralized setting.

6 Conclusion and Work in Progress

The combination of the penetration of Internet into homes, huge computing power at the edge of the network, an exponential growth of user-generated content, a striking need for personalizing Web navigation with respect to search, notification, recommendation, and a call for decentralization to remove the fear of the *Big brother* syndrome and the potential vulnerabilities to attacks of centralized systems, paves the way for a new generation of systems. GOSSPLE should hopefully be one of them. The main originality of GOSSPLE is to make every user responsible for harvesting the Web in a personalized way through the use of efficient gossip-based protocol. Apart from the GOSSPLE challenge that we mentioned above, the challenge of digging out the right tools and scientific backgrounds from as many areas as distributed computing, information retrieval and database is a challenge in itself.

Personalization has been in the air for a while. This has been even more striking as users generate contents. Yet, we are not there yet and combining personalization, security and scalability remains an open track that GOSSPLE tries to fill.

Many challenges need to be tackled in GOSSPLE There are currently three main tracks currently under investigation.

Personalized networks. At the core of Gossple lies the notion of personalized network. GOSSPLE achieves this through gossiping: based on a random peer sampling protocol providing each user with a random subset of other users, GOSSPLE implements a biased sampling protocol that speeds up convergence. Each user periodically contacts a close user, they exchanged their knowledge on the other users and retain the best ones according to a given metric to form the personalized network. Such a protocol enables the quick discovery of related (with respect to a given metric) users in a very large system in a fully distributed manner and with every user storing a small amount of information about the system.

Query expansion in GOSSPLE. In this work, we provide a personalized query expansion mechanism. In the context of a collaborative tagging system ala delicious, Gossple builds, for each user, a personal network of acquaintances through a gossip protocol as explained above. This network is composed of a set of other users that together cover all the interests of the user. This is achieved without revealing the associations between users and their profiles. The information gathered from the personal network is used to create a personalized view of the correlations between tags. This data structure called

the TagMap represents a user-centric personalized view of the relations between tags and is used to expand queries in a meaningful manner. Experimental results conducted on traces crawled from CiteULike, a collaborative tagging system for bibliographic references, and Delicious, show that by storing and exchanging little information between users, the user experience is improved through the query expansion mechanism both with respect to the quality of the results and the number of results obtained. More details can be found in [10].

Top-k processing in GOSSPLE. We are considering decentralized and personalized top-k processing, the protocol is called P3K [5]. It has been shown in [2] that personalizing top-k processing could significantly improve the quality of the results. This was achieved firstly in a centralized way and secondly considering that a social network was known explicitly in advance. We go beyond this approach in P3K. We discover a personal network of acquaintances computing a distance between users based on the similarities observed in their tagging behaviors. In this protocol, we show that using only the information gathered from similar users in a decentralized way, we are able to achieve similar results to those of a centralized approach. We are currently studying a gossip-based alternative to process personalized top-k queries, improving the scalability of the system.

Acknowledgments

I would like to warmly thank all the members of the GOSSPLE team: Xiao Bai, Marin Bertier, Antoine Boutet, Davide Frey, Kevin Huguenin, Vincent Leroy, Afshin Moin, Guang Tan, Christopher Thraves, as well as Rachid Guerraoui who is actively collaborating with us on the project. I also would like to thank Jacques-Henri Jourdan, Fabrice le Fessant and Vivien Quéma for their help.

References

1. Alon, N., Awerbuch, B., Azar, Y., Patt-Shamir, B.: Tell me who i am: An interactive recommendation system. In: ACM Symposium on Parallelism in Algorithms and Architectures (2006)
2. Amer-Yahia, S., Benedikt, M., Lakshmanan, L., Stoyanovich, J.: Efficient network aware search in collaborative tagging sites. In: International conference on Very Large Data Bases (VLDB), vol. 1, pp. 710–721. VLDB Endowment (2008)
3. Amer-Yahia, S., Marlow, C., Yu, C., Stoyanovich, J.: Leveraging tagging to model user interests in del.icio.us. In: AAAI SIP: Social Information (2008)
4. Awerbuch, B., Scheideler, C.: Towards scalable and robust overlay networks. In: International Workshop on Peer-to-Peer Systems (2007)
5. Bai, X., Bertier, M., Guerraoui, R., Kermarrec, A.-M.: Toward personalized peer-to-peer top-k processing. In: International workshop on Social Network Systems (2009)
6. Banaei-Kashani, F., Shahabi, C.: Swam: a family of access methods for similarity-search in peer-to-peer data networks. In: ACM Conference on Information and Knowledge Management (2004)
7. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: International conference on World Wide Web (WWW), pp. 501–510. ACM, New York (2007)

8. Bender, M., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Xavier Parreira, J., Schenkel, R., Weikum, G.: Exploiting social relations for query expansion and result ranking. In: International Conference on Data Engineering Conference (ICDE) Workshops (2008)
9. Bender, M., Crecelius, T., Kacimi, M., Michel, S., Xavier Parreira, J., Weikum, G.: Peer-to-peer information search: Semantic, social, or spiritual? Bulletin of Computer Society Technical Committee on Data Engineering (2007)
10. Bertier, M., Guerraoui, R., Kermarrec, A.-M., Leroy, V.: Toward personalized query expansion. In: International workshop on Social Network Systems, SNS (2009)
11. Sheth Cardoso, A.: J. Semantic Web Services: Theory, Tools and Applications. Springer, Heidelberg (2007)
12. Carman, M., Baillie, M., Crestani, F.: Tag data and personalized information retrieval. In: ACM Workshop on Search in Social Media, SSM (2008)
13. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic analysis of tag similarity measures in collaborative tagging systems. In: Workshop on Ontology Learning and Population (2008)
14. Crespo, A., Molina, H.H.G.: Semantic overlay networks for p2p systems (2002)
15. Eugster, P., Handurukande, S., Guerraoui, R., Kermarrec, A.-M., Kouznetsov, P.: Lightweight probabilistic broadcast. ACM Transaction on Computer Systems 21(4) (November 2003)
16. Eyal, A., Gal, A.: Self organizing semantic topologies in p2p data integration systems. In: International Conference on Data Engineering Conference, ICDE (2009)
17. Fogaras, D., Rácz, B., Csalogány, K., Sarlós, T.: Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. Journal of Internet Mathematics 2(3), 333–358 (2005)
18. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
19. Jelasity, M., Babaoglu, O.: T-Man: Gossip-based overlay toplogy management. In: Brueckner, S.A., Di Marzo Serugendo, G., Hales, D., Zambonelli, F. (eds.) ESOA 2005. LNCS (LNAI), vol. 3910, pp. 1–15. Springer, Heidelberg (2006)
20. Jelasity, M., Montresor, A.: Epidemic-style proactive aggregation in large overlay networks. In: International Conference on Distributed Computing Systems, ICDCS 2004 (2004)
21. Jelasity, M., Voulgaris, S., Guerraoui, R., Kermarrec, A.-M., van Steen, M.: Gossip-based peer sampling. ACM Transactions on Computer Systems (August 2007)
22. Jie, H., Zhang, Y.: Personalized faceted query expansion. In: SIGIR (2006)
23. Kleinberg, J.: The small-world phenomenon: An algorithmic perspective. In: ACM Symposium on Theory of Computing (2000)
24. Lawrence, S.: Context in web search. IEEE Data Engineering Bulletin 23, 25–32 (2000)
25. Le Merrer, E., Kermarrec, A.-M., Massoulié, L.: Peer-to-peer size estimation in large and dynamic networks: a comparative study. In: IEEE International Symposium on High Performance Distributed Computing, HPDC 15 (2006)
26. Milgram, S.: The small-world problem. Psychology Today, 60–67 (1967)
27. Mislove, A., Gummadi, K., Druschel, P.: Exploiting social networks for internet search. In: HotNets. ACM, New York (2006)
28. Monroe, D.: Just for you. Communications of the ACM 52(8) (2009)
29. Morrison, J.: Tagging and searching: Search retrieval effectiveness of folksonomies on the world wide web. Journal of Information Processing and Management (2008) (Corrected Proof) (in press)
30. Niwa, S., Doi, T., Honiden, S.: Web page recommender system based on folksonomy mining for itng 2006 submissions. In: International Conference on Information Technology: New Generations, INTG (2006)

31. Noll, M., Meinel, C.: Web search personalization via social bookmarking and tagging. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 367–380. Springer, Heidelberg (2007)
32. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In: Guerraoui, R. (ed.) Middleware 2001. LNCS, vol. 2218, p. 329. Springer, Heidelberg (2001)
33. Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Xavier Parreira, J., Weikum, G.: Efficient top-k querying over social tagging networks. In: SIGIR (2008)
34. Schenkel, R., Crecelius, T., Kacimi, M., Neumann, T., xavier Parreira, J., Spaniol, M., Weikum, G.: Social wisdom for search and recommandation. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 31(12), 40–49 (2008)
35. Sedmidubsky, J., Barton, S., Dohnal, V., Zezula, P.: Adaptive approximate similarity searching through metric social networks. In: International Conference on Data Engineering Conference, ICDE (2008)
36. Speretta, M., Gauch, S.: Personalized search based on user search histories. In: IEEE/WIC/ACM International Conference on Web Intelligence (2005)
37. Stoica, I., Morris, R., Karger, D., Kaashoek, F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. In: ACM SIGCOMM Conference, San Diego, California (2001)
38. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: International conference on World Wide Web (2004)
39. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: SIGIR (2007)
40. Teevan, J., Dumais, S.T., Horvitz, E.: Characterizing the value of personalizing search. In: SIGIR, pp. 757–758. ACM, New York (2007)
41. Voulgaris, S., Rivière, E., Kermarrec, A.-M., van Steen, M.: Sub-2-sub: Self-organizing content-based publish and subscribe for dynamic and large scale collaborative networks. In: International Workshop on Peer-to-Peer Systems (2006)
42. Voulgaris, S., van Steen, M.: Epidemic-style management of semantic overlays for content-based searching. In: Cunha, J.C., Medeiros, P.D. (eds.) Euro-Par 2005. LNCS, vol. 3648, pp. 1143–1152. Springer, Heidelberg (2005)
43. Wang, Q., Li, R., Chen, L., Lian, J., Tamer Özsu, M.: Speed up semantic search in p2p networks. In: ACM Conference on Information and Knowledge Management (2008)
44. Watts, D.J., Stogatz, S.H.: Collective dynamics of small-world networks. Nature 393, 440–442 (1998)
45. Wong, B., Slivkins, A., Sirer, E.G.: Approximate matching for peer-to-peer overlays with cubit. Technical report, Cornell University, Computing and Information Science Technical Report (2008)
46. Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y.: Exploring folksonomy for personalized search. In: SIGIR, New York, NY, USA (2008)
47. Yildirim, H., Krishnamoorthy, M.S.: A random walk method for alleviating the sparsity problem in collaborative filtering. In: ACM Conference on recommander systems, RecSys (2008)
48. Zanardi, V., Capra, L.: Social ranking: uncovering relevant content using tag-based recommender systems. In: ACM Conference on recommander systems, RecSys (2008)
49. Zhao, S., Du, N., Nauerz, A., Zhang, X., Yuan, Q., Fu, R.: Improved recommendation based on collaborative tagging behaviors. In: International conference on intelligent user interfaces (2008)