The Effect of Introducing Content Price in Distributed Social Networks

Christos Tryfonopoulos
Dept. of Informatics and Telecommunications
University of the Peloponnese
GR22131, Tripoli, Greece
trifon@uop.gr

***Abstract*—Over the last few years a number of distributed social networks with content management capabilities have been introduced both by academia and industry. However, none of these efforts has so far focused on supporting both information retrieval and filtering functionality in a distributed social networking environment. In this work we present a social networking architecture that offers both functionalities –in addition to the usual social interaction tasks– in distributed social networks, outline the associated distributed protocols, and introduce a novel data source selection mechanism for identifying good data sources. This novel data source selection mechanism is designed to take into account a combination of resource selection, predicted publication behaviour, and content cost to improve the selection of information producers by users. To the best of our knowledge our approach, coined AGORA, is the first work to model the price of content and to study its effect on retrieval efficiency and effectiveness in a distributed social network setting. Finally, our work goes beyond modelling by providing proof-of-concept experiments with real-world corpora and social networking data.**

***Index Terms*—distributed social networks, content management, information retrieval/filtering, content price, economic modelling, experimental evaluation**

I. INTRODUCTION

Much information of interest to humans is available today on the Web, making it extremely difficult to stay informed without sifting through enormous amounts of information. In addition, a vast amount of this information is published and shared through social networking sites by users that participate in 'social' activities through the generation, commenting, tagging, liking and sharing vast amounts of digital content. As users engage increasingly more in the usage of social networks, demand for content management capabilities has forced social networks to go beyond the usual social interactions (e.g., like, post, or poke) and offer basic content management functionality. To this end, many social networks adopt the traditional approach of content search (information retrieval - IR) [23], [22], [17], [26], [33], [28]. All these approaches however ignore that the information filtering paradigm may also provide a good alternative of keeping the users informed and at the same time avoiding the information avalanche. In information filtering (IF) –also referred to as publish/subscribe, continuous querying, information push, or information dissemination– users are able to subscribe to information sources and be notified when content of interest is published. This need for content-based push technologies to cope with the information explosion is also stressed by the deployment of tools such as Google Alert and CNN alerts. In an IF scenario, a user posts a subscription (or continuous query) to the system to receive notifications whenever certain events of interest take place (e.g., when a blog post on Winter Olympics becomes available).

However, the increasing usage of social networks gave rise to skepticism about use of centralised services that are able to withhold all uploaded content. Such centralised social networking services are typically owned by private companies and users need to upload their content, thus giving away ownership rights to make it available to others. This allows companies to exploit user data and sell them to advertisers for profit. To circumvent such practises, distributed social networking services –building upon results from the P2P paradigm (e.g., [29], [31], [40])– have been proposed both by academia and industry in the form of distributed social platforms [34], [14], [14], [35] and distributed social content management systems [23], [22], [17], [26], [33], [28]. Although all these approaches offer different types of distributed social networks that allow users to create communities, share content, and send messages, none of them focuses on supporting *information filtering* functionality in such a setup. Moreover, such distributed environments without centralised control, prove an interesting business model for pricing the content delivered by an information producer. In this setup, each information producer (e.g., a news agency, a digital library, or a prolific blogger) might have its own customer base of interested followers/subscribers, and may charge the delivered content by subscription or per item.

In this work, we present AGORA, a *distributed social networking architecture* that allows users to *share, search for* (IR), and *subscribe to* (IF) content in a fully decentralised way, while at the same time maintaining ownership of their content. Such a design is ideal for creating social content marketplaces[1], where users are able to price and distribute their content while at the same time they maintain ownership rights. Content in our setup can be textual, such as status updates/blog posts/tweets, or multimedia, such as photos or videos, appropriately tagged. Our proposed solution offers

[1] AGORA is the Greek word for marketplace

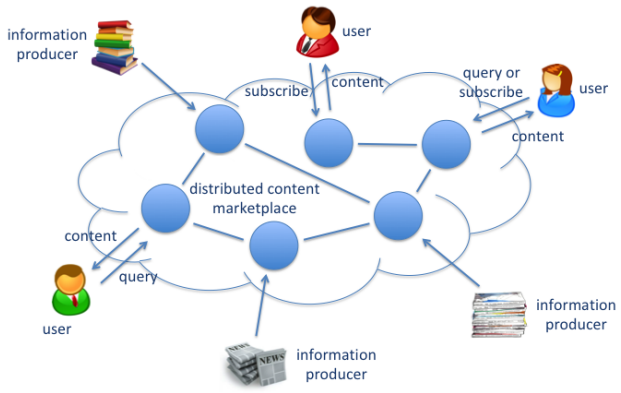

Fig. 1. High-level view of the AGORA architecture.

fundamental social interactions, while emphasising on *content management (IR and IF)* and economic aspects such as the *price/quality tradeoff* of content. To the best of our knowledge, this is the first approach that aims at combining IR and IF in a social networking context, while taking into account aspects of economic modelling. In the light of the above, the contributions presented in this work are the following:

- We propose a social networking architecture that offers content management functionality in terms of IR and IF, in addition to the usual social interaction tasks typically supported in such scenarios. This is the first approach in the literature to offer both functionalities.
- We present the *distributed protocols* and *services* that regulate node interactions, provide details on the distributed IR and IF, and outline the different friendship facets introduced in the architecture.
- We devise a novel method to rank information producers according to (i) the content already published, (ii) the expected future publishing behaviour, and (iii) the price announced by the information producer. This method allows us to achieve high recall with low cost (by searching at/subscribing to a small number of information producers).
- We study the effect of content price in our setup and experimentally demonstrate that it is a key element on the delivered content quality. Our modelling utilises concepts such as correlation between the quality/expertise of the information producer, demand-driven price rates, and cost of resources.

Figure 1 shows a high-level view of the envisioned distributed architecture with different types of information producers and users with varying information needs (IR, IF or both).

The rest of the paper is organised as follows. Section II overviews related research, while Section III discusses the AGORA architecture and the associated distributed protocols. Subsequently, Section IV highlights the economic and qualitative aspects of AGORA and presents the price/quality tradeoff, while Section V experimentally demonstrates the effect of cost in retrieval and filtering quality and compares the system effectiveness with/without monetary flow. Finally, Section VI discusses high-level conclusions, open issues, and possible extensions.

II. RELATED WORK

In this section, we discuss related work in the context of systems that are designed with content management in mind, and platforms that provide distributed social networks.

A. Distributed Social Platforms

All available social networks (e.g., Facebook, LinkedIn, Elgg) are currently based on centralised solutions both for storing and managing of content, which set scalability limitations on the system and reduce fault-tolerance. Industry has already detected these drawbacks and has lately turned into solutions that diverge from the centralised model of the existing systems by developing platforms, such as Diaspora, KrawlerX and OpenSocial, that provide APIs to support application hosting in remote application servers, owned and managed by the application providers. In a similar spirit, a strand of research work also moved towards hierarchical organisations for supporting distributed social networking. The distributed social systems SuperNova [34] and Scope [24] are based on a two-tier architecture, where nodes with higher computing capability become super-nodes and form an overlay to provide distributed data management of the P2P social network. Client nodes connect to super-nodes and rely on them for bootstrapping, sharing their content, and accessing the shared information. Although all of the platforms and system schemes propose the decentralisation of the social services, one of the main issues of the centralised architectures persists: the existence of a single point where user information is collected and may be exploited.

To alleviate the above disadvantage, distributed platforms for social online networks based on the P2P paradigm were proposed [27], [1], [14]. LifeSocial.KOM [14] is a plugin-based extendible social platform that provides secure communication and user-based data access control, and integrates a monitoring component that allows users and operators to observe the quality of the distributed system. Similar efforts aimed at spontaneous social networking; they include proposals for distributed social services in resource constrained devices (like tablets or smartphones) [35], [8], or in environments with no infrastructure guarantees (e.g., high-attendance events) [24]. All these approaches offer different types of distributed social platforms that allow users to create communities, share content, and send messages, but do not emphasise expressive content search mechanisms.

Finally, other approaches in distributed social networking emphasise on delivering innovative and competitive services; SCIMS [19] relies on an ontology-based model for managing social relationships and status, the work in [37] aims at personalising search results based on user context and friendship relations, while Gemstone [36] targets data availability in the absence of the data owner. To achieve this, a replica storage

scheme based on social relationships, online patterns of nodes, and user experiences is utilised.

B. Distributed Social Data Management

Our work fits mainly into the area of content management in distributed social networks and is inspired by previous approaches on Semantic Overlay Networks (SONs) [29], [31], [40] and based on works that emphasise on distributed content location in social networks. Works like the eXO [23] and SoNet [22] systems are, similarly to AGORA, inspired by the P2P paradigm to provide content location and management services on large-scale decentralised social networks. To do so, the authors rely on a structured overlay and exploit the accurate location mechanisms, but de-emphasise node autonomy. Contrary to these approaches, AGORA employs a loose component architecture and introduces a new type of social relations between nodes: the semantic closeness of content. In this way, nodes that are similar in terms of content, create emergent groups likewise to the creation of social relations. Our work shares ideas with the SocialCDN system [17], where social caches (links among friends) are introduced as a way to alleviate the network traffic and optimise data dissemination (mainly by social updates). In [17], social cache selection is formulated as the neighbour-dominating set problem and a family of algorithms is proposed and evaluated. Contrary to AGORA, where the emphasis is on efficiently supporting expressive content retrieval in the social paradigm, the emphasis on SocialCDN is on the reduction on network traffic to facilitate fundamental social interactions.

The loose component architecture and the emphasis on node autonomy of [26] resemble the architectural design of AGORA, where an unstructured overlay network of nodes is utilised to support the distributed social infrastructure. However, the focus of [26] is on the design of gossip protocols for efficiently disseminating profile updates to all interested users and does not put any attention to the problem of content search and management.

Furthermore, the concept of creating and maintaining social connections in distributed infrastructures is affined with the problem of distributed data management in P2P networks. In SONs (e.g., [40], [6]), "social" connections between the peers (e.g., similarity of content, pattern, or distance in a physical level) are exploited to direct the search to nodes with relevant data (e.g., as in [32] that studies query routing strategies based on "social" relationships). Other works on SONs (e.g., [31]) focus more on the organisation of P2P networks as small-world networks, where peers self-organise in groups of similar interests to facilitate message-efficient query answering. Our work on AGORA borrows concepts and ideas from research on SONs and extends them for facilitating efficient and effective data management in a social network setting. We suggest that SONs offer the most promising architectural solution inspired from the P2P paradigm; it is a perfect fit for a distributed social networking scenario providing high decentralisation, high node autonomy, support for emergent semantic and social structures, and effective object location mechanisms. Contrary, DHT-based architectures [23], [28] ignore node autonomy (by enforcing deterministic key/content placement) and emergent structures (by enforcing network structure).

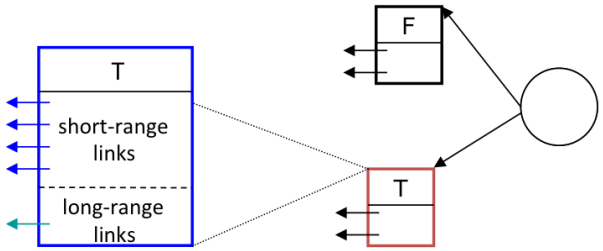

Fig. 2. The thematic and friend index of a node.

Finally, a large number of research in the domain of distributed social networks consists of studies on system security [28], [7], user privacy [28], [7], [16], [15], [41], distributed access control [2], [4], and authentication mechanisms [2]. Clearly, security issues are also relevant in our design and the AGORA system could benefit by adopting approaches like [2] or [15] that enforce user privacy and access control. However, the problem of security is orthogonal to our design and is not further analysed as it is not the emphasis of this work.

III. System Overview

In this section, we provide an overview of the architectural components of the proposed system and provide the distributed protocols executed locally by users and information producers.

A. Architecture

We consider a distributed social network, where each user, characterised by its interests, is connected to friends and information producers with similar interests. The interests of a user are identified automatically, i.e., by applying clustering on its local content repository. The network nodes use a heartbeat protocol that runs continuously and aims at identifying new information producers based on the likelihood to have similar interests with the node at hand. Each user maintains two *routing indices* holding information for *social friends* and *thematic friends*. Social friend links correspond to the social relationship aspect of the network, while thematic friend links are of two types: short-range links (i.e., links to information producers with similar interests) and long-range links (i.e., links to information producers with dissimilar interests to maintain connectivity between different clusters in the system). The reorganisation procedure is executed locally by each node and aims at bringing together users and information producers with similar interests that are likely to share/search for/subscribe to content.

B. Joining Agora

When a user node connects to the network, its interests are automatically derived by its local content. For each interest, the node maintains a thematic index (T) containing the contact details and interest descriptions of information producers. These links form the thematic neighbourhood of the node;

the links contained in T are refined accordingly by using the rewiring service described below. Furthermore, each node maintains a friend index (F) containing the contact details and interest descriptions of the social neighbourhood of the node, comprised of explicitly declared friends in the network. Figure 2 shows the F and T for an arbitrary node.

C. Locating information producers

This service is applied to locate new information producers by establishing new connections and discarding old ones. Each node may initiate this procedure by computing a scoring function that combines the quality and price of all information producers in its thematic index (T). If the computed score is greater than a threshold then the node does not need to take any further action, since it is already aware of information producers that match its needs and budget. Otherwise, the node initiates a process to identify new information producers by forwarding a message in the network –bounded by a time-to-live (TTL) mechanism– using the thematic and social connections and collecting the interests of other information producers.

The issued message is forwarded with equal probability to (i) a number of randomly chosen entries contained in a node's T, (ii) a number of randomly chosen entries contained in a node's F, or (iii) the most similar nodes to the message initiator, found in either T or F. The rationale of applying either of the forwarding strategies is that the message initiator should be able to reach information producers both directly (through other similar nodes), but also indirectly (through propagation of the message through non-similar nodes). Each node that receives the message adds its interest in it, reduces TTL by one, and forwards it in the same manner. When the TTL of the message reaches zero, the message containing the contact info and interests of all information producers that received the message is sent back to its initiator. To speed up the process, every intermediate node receiving the message may utilise the information in it to refine its thematic connections.

D. Subscribing for content at information producers

Queries/subscriptions are issued as free text or keywords under the Vector Space Model and are formulated as term vectors. The user subscribing for specific content forwards a message in the network with a TTL using both its social and thematic connections. The issued message is forwarded both to (i) friends that have interests similar to the query and are contained in F and (ii) a small number of information producers contained in the T chosen as described below. Initially, the message initiator compares the user subscription against its interests and, if similar, the message is forwarded to all of its short-range links, i.e., the message is *broadcasted* to the node's neighbourhood (*explosion*). Otherwise, the message is forwarded to a small fixed number of nodes that have the highest similarity to the subscription (*fixed forwarding*). The combination of the two routing strategies is referred to in the literature as the *fireworks* technique [29]. All the nodes receiving the message reduce TTL by one and apply the same forwarding technique; the message is not forwarded further in the network when TTL reaches zero. Additionally to forwarding, every node receiving a message compares the subscription against its identified interests and, if similar, stores it in its local continuous query data structures to match it against future publications.

Content is kept locally at each information producers in the spirit of [42]. This lack of publication dissemination is a design decision that offers increased scalability (by trading recall) and complies with the content marketplace paradigm we aim for. Thus, only the nodes indexing a subscription can notify the interested user, although other information producers may also publish relevant content. When an information producer wants to publish new content to the network, it matches it against its local continuous query database to decide which continuous queries match it and thus, which user node should be notified. Then, the information producers delivers a notification for each continuous query by sending to the query initiator a pointer to the matching content; if the user is not online, the provider stores the message and delivers it upon user reconnection. Notice that since the information producer maintains the content ownership, it is now able to charge the user at the announced content price.

E. Searching for content at information producers

A user issuing content search forwards a message in the network following a mechanism similar to the one described in the previous section. Additionally to query forwarding, every information producer receiving a query message compares it against the identified interests and, if similar, matches it against the locally stored content. Subsequently, pointers to the matching content are sent to the query initiator (along with the announced price for the content). Subsequently answers from all information producers are ranked by a combination of price and similarity to the query (discussed in the next section) and the list is presented to the user.

IV. Ranking of information producers

To select which information producer will be monitored, the protocols described in the previous section use a scoring function to rank information producers. In this section, we quantify these concepts and give the rationale between our choices.

A. Quality vs Price

The ranking strategy for the information producers is a critical component since it allows users to locate relevant information and information producers to maximise their revenue. Empirical studies have shown that price and quality are the two key determinants of the consumer's choice to buy or not a product [10]. Thus, to make an informed selection on the information producers, a user has to rank them based on a blend (denoted by tunable parameter β) of content quality and content price both in a retrieval and a filtering scenario. This combination describes the benefit/cost ratio for the content and

allows the user to assign a score to every information producer. To this end, the score for each information producer is given by:

$$score(p,i) = \beta \cdot price(p,i) - (1-\beta) \cdot quality(p,i) \quad (1)$$

Here, p denotes an information producer, q denotes an *information need* (in the form of a query or a subscription), $price(p,i)$ refers to the announced price an information producer provides for the provided content, and $quality(p,i)$ denotes how relevant p is to i. Information producers compute the price based on the demand and according to popularity of themselves and that of their content. Price and quality have the same range of values to allow for their combination, while price is typically recomputed whenever the popularity of the information producer changes. In Section V we study the price in different scenarios, and show the effect on the quality of retrieved content (i.e., recall) when the price choice is (i) random, (ii) strongly correlated with quality, and (iii) partially correlated with quality.

B. Content price specifics

In this section, we analyse the economic modelling of AGORA and review the basic assumptions and expectations from such a modelling. Usefulness of the information goods received by a subscriber is a qualitative criterion, that is difficult to model. In AGORA, we model usefulness by matching interests, i.e., by assuming that all received content relevant to the information need are useful to the subscriber, and do not discuss issues such as novelty, coverage of the field, or user effort. In our modelling, after a user acquires a history of transactions with certain information producers, develops an affect for some of them. Affect can be modelled in various ways, depending on the task at hand, and can be either positive or negative [11]. In AGORA, a user does not a priori know the quality of the information goods, but uses the affect developed from previous transactions to approximate it. Subsequently, he compares the values of information quality to the expected values and update its affect [21].

The costs in AGORA are results of actions [18], such as transactions, network communication, and use of common infrastructure. Information producers try to maximise their revenue while minimising expenses that occur due to their actions. Contrary users try to maximise the utility of the received content, while minimising expenses that occur due to actions. In general, the content market in AGORA is not a pure competitive market in the sense of e.g., [39], since users do not know in advance the exact content quality they are buying. AGORA resembles the modelling of a team of sales people [38], where stakeholders try to collaborate with others in order to get their expertise for a (cross/up) sale. After deciding who to collaborate with, it is possible to model the gap between the initial expectations and the actual actions. In [12], it is shown that this gap is smaller in a competitive relationship compared to that of a cooperative relationship. As in many cooperative environments each stakeholder usually retains its connections with the others, while also being free to explore new mutually beneficial connections. This is in line with the notion of friendship connections in any social network.

The main goal of this modelling is to study the influence of the cost component on the quality of received content, study the interactions between users and information producers, and gain insights about the overall behaviour of this prototype content market.

C. Content quality specifics

The quality of the content is a difficult thing to model as it cannot be known prior to acquiring it. Hence the best option is to assess the quality of the information producer. To do so, one has to take into account the dual capacity of information producers in AGORA: to answer one-time content requests (IR) and satisfy long-term information needs (IF). To this end, the quality of an information producer has to be based on (i) the quality of already published content (since this depicts its ability to satisfy IR tasks from users) and (ii) a predicted quality of the content to be published in the future (since this depicts its ability to satisfy IF tasks from users). The necessity of both facets is better illustrated in the case of an information producer that provides content in the form of technical articles and news items; articles have a long shelf-life and are good candidates for recurrent sales, while news items have an extremely short shelf-life and after some time users loose interest in them.

1. Quality of published content. The quality of the already published content is known as the *resource selection problem* in the areas of information retrieval and databases. Hence, a number of standard resource selection algorithms such as tf-idf based methods, CORI, or language models (see [30] for a nice overview) have been proposed. We use a standard tf-idf based component that combines accuracy of modelling and ease of implementation. Thus, in our setup, quality of published content is given by:

$$\sum_{t \in i} [0.5 \cdot \log(df_{p,t}) + 0.5 \cdot \log\left(tf_{p,t}^{max}\right)] \quad (2)$$

Our approach uses the standard IR constructs of document frequency (df) and maximum term frequency (tf^{max}) [25], while a balanced combination of these two metrics (0.5 in the above equation) is used to equally emphasise importance of df and tf^{max} according to [30], [43].

2. Predicted quality of content to be published. To predict the quality of the content to be published, we model IR statistics per information producer as time-series data and use statistical analysis tools [5] to predict future values based on past observations. Such techniques take into account assumptions about the internal structure of the time series like trends and seasonality and tend to emphasise recent observations. Since seasonality requires long-term statistics that are infeasible to maintain (e.g., several years of data to observe seasonality in the publication of Christmas content), we resort to *double exponential smoothing* techniques [5] as our prediction mechanism, since it requires a small amount of

data, emphasises recent over older data observations, and allows for trend identification (useful for news items or trending topics). Thus, in our setup, predicted quality of content to be published is given by:

$$\sum_{t \in i} \log\left(\delta(df^*_{p,t}) + \log\left(\delta(cs^*_p) + 1\right) + 1\right) \quad (3)$$

In the formula above, function $\delta(df^*_{p,t})$ stands for the difference between the predicted and the last document frequency (df) observed, and $\delta(cs^*)$ is the difference in the collection size of information producer p reflecting its overall expected future publishing activity. These values are calculated for all terms t contained in the user subscription. In this way we model two aspects of the information producer's behaviour: (i) its potential to publish relevant documents in the future (reflected by $\delta(df^*_{p,t})$) and (ii) its overall expected future publishing activity (reflected by $\delta(cs^*)$). Notice also that, in the above formula, the publication of relevant documents (i.e., $\delta(df^*_{p,t})$) is more emphasised than the publication rate ($delta(cs^*)$) due to the nesting of the $\log$ functions. The addition of 1 in the $\log$ functions is used to yield positive predictions and avoid $\log(0)$.

3. Putting it all together. Given Formulas 2 and 3, the overall quality of an information producer p with respect to an information need i is given by:

$$quality(p,i) = \sum_{t \in i}[(0.5 \cdot \log\left(df_{p,t}\right) + 0.5 \cdot \log\left(tf^{max}_{p,t}\right)) + \log\left(\delta(df^*_{p,t}) + \log\left(\delta(cs^*_p) + 1\right) + 1\right)] \quad (4)$$

V. Experimental Evaluation

In this section, we present our findings from the introduction of content cost, and how it affects the effectiveness of the system. We study the behaviour of AGORA using different scenarios, while varying the correlation between price, quality and customer demand.

A. Experimental Setup and Metrics

We used a real-life document collection containing $2M$ documents with a vocabulary of about $600K$ terms from a focused Web crawl. The documents were categorised in ten different categories by content, with the smallest category having $67K$ documents and the largest one of $325K$. In all experiments, the network consists of $1,000$ information producers and $1M$ users. Each information producer started with a database of 300 documents initialised with 15% random category, 10% non-categorised, and 75% single category documents, resulting in 100 specialised information producers for each category. We artificially constructed information needs for the users (in the form of queries and subscriptions with multiple terms) using the document corpus and by selecting terms that are strong representatives of a document category (i.e., a frequent term in documents of one category and infrequent in documents of the other categories). The simulation was done in rounds.

We introduced budget constraints on a per user basis; thus, each user was able to follow only a few selected information producers that were chosen according to his own information needs and ranked using the formulas discussed in Section IV. For deciding the per user budget, we relied on studies about budget distribution and spending for a variety of cases, ranging from family budgets to consumer budgets [20], [9]. The main conclusions drawn from these studies are that (i) budget distribution follows a power law, with a small percentage of families/consumers having a high (yearly) budget, and a large percentage of the families being in the (long) tail of the distribution, with a low budget, and (ii) the percentage of the income spend on content does not vary with the budget. According to [3] and the above remarks, we divided the users into three classes: low, average and high budget users with a population of $600K$ (or 60% of total users), $300K$ (or 30% of total users) and $100K$ (or 10% of total users) respectively. Subsequently, we experimentally computed a budget that would allow the users to subscribe to the top-100 (i.e., 10%) information producers, and allowed the low budget users to have 60%, the medium budget users to have 80% and the high budget users to have 120% of this ideal budget.

To measure the effect of content cost in quality (with and without monetary flow) we utilise the following metrics:

- *Messages.* We measure the message traffic per action in AGORA.
- *Recall.* We measure *average recall* over all rounds by computing the *ratio* of the total number of retrieved content to the total number of relevant content for all users.
- *Ranking.* We use an extension of Spearman's footrule distance to compare rankings of information producers calculated by users. This metric allows us to compare two different information producer rankings by calculating the distance between the elements in two ranking lists. In our implementation if an element from list A is not present in list B, it is considered as being in the last available position in B.

Finally, to assist the reproducibility of our results we plan to publicly release our code in an appropriate repository after the publication of the work.

B. Varying the price-quality correlation

In this experiment, we aimed at observing the behaviour of AGORA when varying the correlation $0 \leq \kappa \leq 1$ between the price and the quality of an information producer; $k = 1$ means that the price an information producer charges for content is fully correlated to the quality of that producer. Notice that quality in AGORA is easily calculated/forecasted using Equation 4 and in this case users know that the content from this information producer will be expensive but relevant (i.e., useful to them). The other extreme is when $k = 0$, where prices have no correlation to the quality of the information producer and are chosen randomly. For the rest of the cases, i.e., $0 < \kappa < 1$, the correlation is modelled as the likelihood that an information producer sells $\kappa\%$ of content underpriced (up to 20% of the initial value), and $1 - \kappa$ overpriced (up to 20% of the initial value).

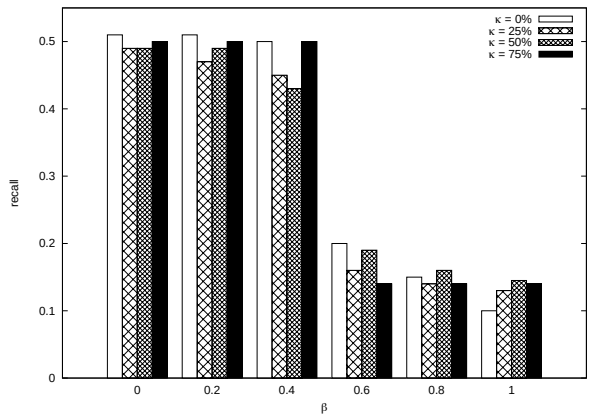

Fig. 3. Recall against β.

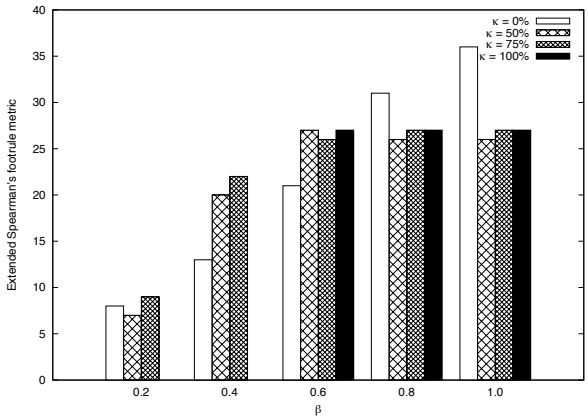

Fig. 5. Difference in information producer ranking against β.

Figure 3 presents the achieved recall for different values of β and κ, where we notice that the introduction of price for content reduces the observed recall (notice that recall has the highest values for $\beta = 0$, i.e., no pricing is involved in information producer ranking). This is an important result, showing that when information producers charge for content, consumers trade quality for cheaper options. Notice also that for every κ value tested recall is retained high, as long as it plays the most important role in the ranking ($\beta < 50\%$). This was also expected, since when price is of importance, consumers will choose cheaper information producers, leading to a reduction in the observed recall. Additionally, when the price is the only ranking criterion for users ($\beta = 1$), recall is close to that of a random choice of information producers (remember that users monitor only 10% of information producers in the system).

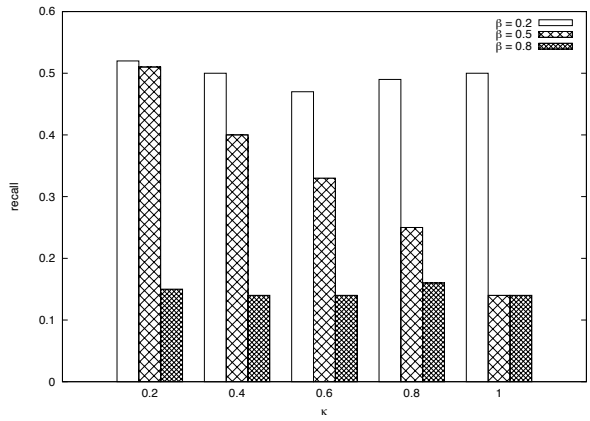

Fig. 4. Recall against κ.

Figure 4 shows how recall is affected for three different values of β, when κ increases. When one of the two components becomes dominant in the ranking function, it outweighs the effect of the other. This is in line with our expectations, since price dominates the ranking function and quality is sacrificed to reduce costs. Finally, notice that consistent recall between different values of κ is an effect of the modelling of AGORA as a closed system where monetary flow is limited by the budgets of the users and no new wealth is produced.

Figure 5 shows how differently users rank information producers for varying values of β and κ. The difference in the ranking of information producers is measured using an extension of Spearman's footrule metric. To produce a point in the graph we compare the ranked lists of information producers for each pair of users for and each value of β in the x-axis and average the results. Notice that the case of $\beta = 0$ is omitted as the extended Spearman's footrule metric is always 0 since the lists compared are identical (users take into account only the quality to rank information producers). When β increases (i.e., price becomes more important in the ranking process) Spearman's metric increases too, as information producers with high quality get lower positions in the ranking, while information producers with lower quality (but cheaper) are ranked high. Finally, when the price of an information producer is not associated with its quality (random price setup or $\kappa = 100\%$), there are big differences in the ranking of information producers (especially in the cases where price matters more – the leftmost points in the graph) due to the introduced randomness.

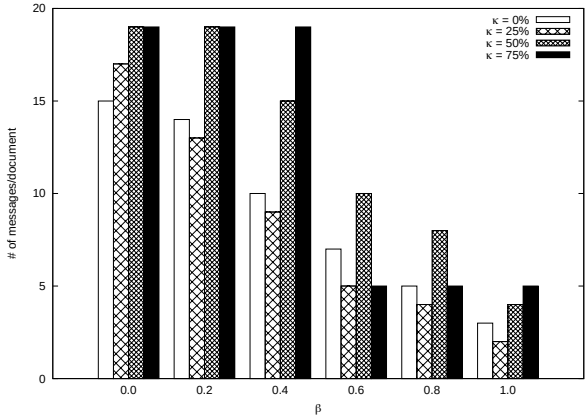

Fig. 6. Message traffic against β.

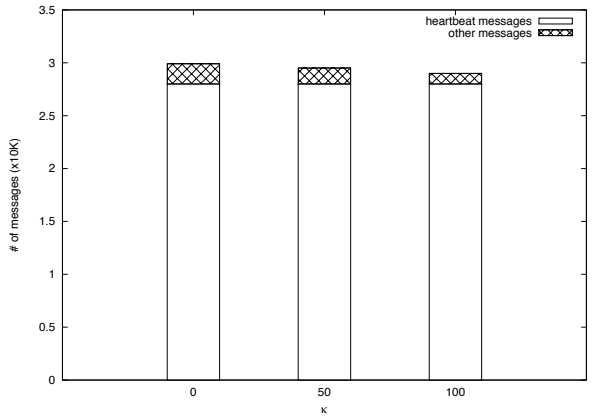

Fig. 7. Message traffic against κ.

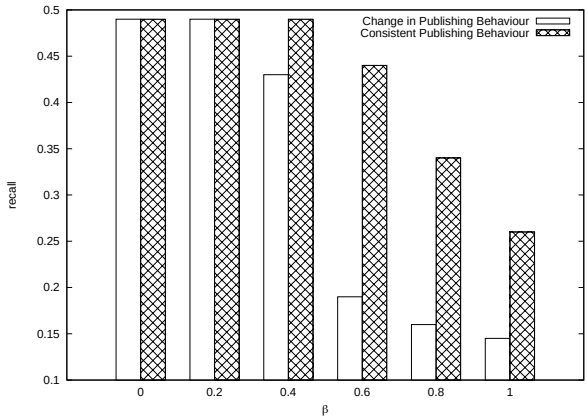

Fig. 8. Recall for different scenarios against β.

C. System Performance

In the first series of experiments we targeted the performance of AGORA in terms of message traffic. Figure 6 shows that the message traffic per user incurred in AGORA is reducing as β increases. This happens because users utilise the price component to rank information producers, thus choosing those of lower price and poor quality. This reduces the overall message number as less content is exchanged in the system due to users subscribing to non-expert information producers; our observations here are consistent with those in Section V-B that correlate recall and β.

Figure 7 demonstrates the total amount of traffic observed in AGORA and how this traffic is split in the various message categories, as the price-quality correlation is varied. As expected the heart-beat protocol messages dominate the messaging load, as necessary messages with information producer statistics and prices are disseminated in the system. Notice also that these messages are not affected by price-quality correlation, since the information producers have to update their publication statistics and prices, regardless of their customer base. Finally, notice that the number of messages for querying for/subscribing to/receiving content are affected as κ increases since information producers that are of high quality widen their customer base with more users.

D. Varying the behaviour of information producers

In this section we look into recall and how this is affected by two very different information producer behaviours: *topic specialisation* and *topic shift*. In topic specialisation, information producers disregard market conditions and maintain their topic specialisation even if this results in lower revenues. Contrary in topic shift, information producers may alter their specialisation topic over time based on changes in user demand, revenue and market conditions. In the latter case, an information producer initially publishes content from one category, and some rounds later may decide to switch to a different category to simulate changes in portfolio or a different business strategy.

Figure 8 shows the observed recall for both scenarios and different values of β. The most important observations in this graph are (i) the drop in recall for both scenarios as β increases and (ii) the higher recall values for the case of topic specialisation. The reason for the first observation is the shift of users to cheaper information producers due to the importance of content price in the ranking function. Additionally, the reason for the second observation is that the build up of expertise by the information producers reflects on the higher quality values, which in turn leads to ranking quality information producers higher. Contrary, when information producers shift their topic, users are not able to correlate price and quality and thus, select information producers of poor quality (hence the drop in recall).

VI. CONCLUSIONS AND FUTURE WORK

In this work, we proposed AGORA, a social networking architecture and the related distributed protocols to facilitate distributed content management in the form of information retrieval and information filtering. In addition, we introduced a novel selection mechanism for information producers that allows users to rank them according to (i) their expertise, (ii) their predicted future content publications, and (iii) the price of content they deliver. The experimental evaluation of our proposal demonstrated the behaviour of such a content marketplace in terms of price/quality tradeoffs, recall and message traffic. To the best of our knowledge, these are the first results that connect recall and message traffic in a distributed social network with the content cost, and put economic modelling at the heart of system design. The most important outcome of our study is that price should participate in the ranking of information producers less than a quarter of the total score to avoid the delivery of irrelevant content to users. We also showed that the introduction of content price improves system scalability by reducing message traffic and imposing a reasonable use of resources to stakeholders. Overall, introducing a monetary value for the production and dissemination of content in a distributed social network proves an interesting business model that conserves resources and improves scalability. However, this should be executed in a careful fashion to avoid user dissatisfaction that may rise from

the content cost itself, and the reduced access to relevant information.

Future directions of research include more extensive experimentation (using a grid service for vast-scale experiments and analytics), more detailed economic modelling (e.g., model AGORA as an open system, perform monetary flow monitoring/analysis, incorporate agent-style BDI modelling [13]), and implementation of a prototype system.

ACKNOWLEDGEMENTS

The author would like to thank the anonymous reviewers for their comments and suggestions on improving the quality and presentation of the work.

REFERENCES

[1] S. Abbas, J. Pouwelse, D. Epema, and H. Sips. A Gossip-Based Distributed Social Networking System. In *WETICE*, 2009.

[2] M. Backes, M. Maffei, and K. Pecina. A Security API for Distributed Social Networks. In *NDSS*, 2011.

[3] L. P. Breker. A survey of network pricing schemes. In *Theoretical Computer Science*, 1996.

[4] S. Buchegger, D. Schioberg, L.-H. Vu, and A. Datta. PeerSoN: P2P social networking: early experiences and insights. In *SNS*, 2009.

[5] C. Chatfield. *The Analysis of Time Series - An Introduction. CRC Press 2004*.

[6] A. Crespo and H. Garcia-Molina. Routing indices for peer-to-peer systems. In *Proceedings of 22nd IEEE International ICDCS Conference*, pages 23–32, Vienna, Austria, July 2002.

[7] L. A. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *IEEE Communications Magazine*, 2009.

[8] A. de Spindler, M. Grossniklaus, and M. C. Norrie. Development Framework for Mobile Social Applications. In *CAiSE*, 2009.

[9] J. B. DeLong. Six Families Budget Their Money. In *Lecture notes for American Economic History, University of California at Berkeley*, 2008.

[10] A. Dumrogsiri, M. Fan, A. Jain, and K. Moinzadeh. A supply chain model with direct and retail channels. In *European Joumal of Operational Research*, 2008.

[11] M. Faig and B. Jerez. Inflation, Prices, and Information and Competitive Search. *Journal of Macroeconomics*, 2006.

[12] L. Forker and P. Stannack. Cooperation versus Competition: do buyers and suppliers really see eye-to-eye? In *European Journal of Purchasing and Supply Management*, 2000.

[13] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. The Belief-Desire-Intention Model of Agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages*, 1999.

[14] K. Graffi, C. Gross, P. Mukherjee, A. Kovacevic, and R. Steinmetz. LifeSocial.KOM: A P2P-Based Platform for Secure Online Social Networks. In *P2P*, 2010.

[15] B. Greschbach, G. Kreitz, and S. Buchegger. The devil is in the metadata - New privacy challenges in Decentralised Online Social Networks. In *PerCom Workshops*, 2012.

[16] F. Gunther, M. Manulis, and T. Strufe. Key management in distributed online social networks. In *WOWMOM*, 2011.

[17] L. Han, M. Punceva, B. Nath, S. Muthukrishnan, and L. Iftode. SocialCDN: Caching techniques for distributed social networks. In *P2P*, 2012.

[22] G. Liu, H. Shen, and L. Ward. An efficient and trustworthy P2P and social network integrated file sharing system. In *P2P*, 2012.

[18] B. Johansson and H. Persson. Self-organised adjustments in a market with price-setting firms. In *Chaos, Solitons and Fractals*, 2003.

[19] M. A. Kabir, J. Han, J. Yu, and A. Colman. SCIMS: A Social Context Information Management System for Socially-Aware Applications. In *CAiSE*, 2012.

[20] B. B. A. Kennickell and K. Moore. Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances, 2006.

[21] C. Li, M. Singh, and K. Sycara. A Dynamic Pricing Mechanism for P2P Referral Systems. In *AAMAS*, 2004.

[23] A. Loupasakis, N. Ntarmos, and P. Triantafillou. eXO: Decentralized Autonomous Scalable Social Networking. In *CIDR*, 2011.

[24] M. Mani, A.-M. Nguyen, and N. Crespi. SCOPE: A prototype for spontaneous P2P social networking. In *PerCom Workshops*, 2010.

[25] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[26] G. Mega, A. Montresor, and G. Picco. Efficient dissemination in decentralized social networks. In *P2P*, 2011.

[27] J. Mitchell-Wong, S. Goh, M. Chhetri, R. Kowalczyk, and Q. Vo. A Framework for Open, Distributed and Self-Managed Social Platforms. In *VECN*, 2008.

[28] R. Narendula, T. Papaioannou, and K. Aberer. My3: A highly-available P2P-based online social network. In *P2P*, 2011.

[29] C. H. Ng, K. C. Sia, and C. H. Chang. Advanced Peer Clustering and Firework Query Model in the Peer-to-Peer Network. In *WWW*, 2002.

[30] H. Nottelmann and N. Fuhr. Evaluating Different Methods of Estimating Retrieval Quality for Resource Selection. In *SIGIR*, 2003.

[31] P. Raftopoulou and E. Petrakis. iCluster: a Self-Organising Overlay Network for P2P Information Retrieval. In *ECIR*, 2008.

[32] P. Raftopoulou, E. Petrakis, and C. Tryfonopoulos. Rewiring strategies for semantic overlay networks. In *DPD*, 2009.

[33] P. Raftopoulou, C. Tryfonopoulos, E. Petrakis, and N. Zevlis. DS4: Introducing Semantic Friendship in Distributed Social Networks. In *CoopIS*, 2013.

[34] R. Sharma and A. Datta. SuperNova: Super-peers based architecture for decentralized online social networks. In *COMSNETS*, 2012.

[35] P. Stuedi, I. Mohomed, M. Balakrishnan, Z. Mao, V. Ramasubramanian, D. Terry, and T. Wobber. Contrail: Enabling Decentralized Social Networks on Smartphones. In *Middleware*, 2011.

[36] F. Tegeler, D. Koll, and X. Fu. Gemstone: Empowering Decentralized Social Networking with High Data Availability. In *GLOBECOM*, 2011.

[37] B. Upadhyaya and E. Choi. Social Overlay: P2P Infrastructure for Social Networks. In *NCM*, 2009.

[38] T. Üstüner and D. Godes. Better Sales Networks. In *Harvard Business Review*, 2006.

[39] H. R. Varian. Pricing Information Goods. In *Research Libraries Group Symposium, Harvard Law School*, 1995.

[40] S. Voulgaris, M. van Steen, and K. Iwanicki. Proactive Gossip-based Management of Semantic Overlay Networks. *CCPE*, 19(17), 2007.

[41] M. Xue, B. Carminati, and E. Ferrari. P3D - Privacy-Preserving Path Discovery in Decentralized Online Social Networks. In *COMPSAC*, 2011.

[42] C. Zimmer, C. Tryfonopoulos, K. Berberich, M. Koubarakis, and G. Weikum. Approximate Information Filtering in Peer-to-Peer Networks. In *WISE*, 2008.

[43] C. Zimmer, C. Tryfonopoulos, K. Berberich, G. Weikum, and M. Koubarakis. Node Behavior Prediction for Large-Scale Approximate Information Filtering. In *LSDS-IR @ SIGIR*, 2007.