University of Zurich[UZH]

Mengia Zollinger
Cosmin Basca
Abraham Bernstein

Market-based SPARQL brokerage with MaTriX: towards a mechanism for economic welfare growth and incentives for free data provision in the Web of Data

TECHNICAL REPORT – No. IFI-2013.04

June 2013

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

Market-based SPARQL brokerage with MaTriX: towards a mechanism for economic welfare growth and incentives for free data provision in the Web of Data

Mengia Zollinger[1], Cosmin Basca[1], Abraham Bernstein[1]

Dynamic and Distributed Information Systems, Department of Informatics,
University of Zurich, Zurich, Switzerland
{lastname}@ifi.uzh.ch

Abstract. The exponential growth of the Web of Linked Data (WoD) has so far primarily been funded using subsidies, where new datasets are financed through public funding or via research programs. Relying on (public) subsidies, however, may eventually limit the growth of the WoD, focus on areas decided by committee rather than true demand, and could hamper data quality due to the lack of clear incentives to maintain high quality standards.

In this paper we propose a market-based SPARQL broker over a heterogenous, federated WoD as a economically viable growth option. Similar to others, we associate each query with a given (potentially zero) budget and a minimal results-set quality constraint. The SPARQL broker then employs auction mechanisms to find a desirable set of data providers that jointly deliver the results. We evaluate our market-based SPARQL broker called `MaTriX` using a simulation. Our results show that a mixture of free and commercial providers actually provide superior performance in terms of consumer surplus, producer profit, total welfare, and recall whilst being incentive compatible with the provision of high-quality results. We even found that the increase of profit in the mixed situation may entice commercial providers to subsidize free providers directly.

1 Introduction

The Web was able to entice people to publish a large amount of unstructured data online. Its growth was fueled by the human desires to promulgate one's thoughts, to become famous, to communicate with each other, and/or earn money paired with the universal accessibility provided by browsers and search engines.

The Web of Data (WoD) is also growing fast. In past years it has doubled (in

terms of numbers of RDF triples asserted) each year to an estimated 31 billion triples.[1] Borrowing from the traditional Web's characteristics the WoD is inherently a decentralized repository of knowledge, where many data tenants publish and manage interlinked RDF datasets wether indexed or not. In contrast to the traditional Web the vast majority of datasets are freely available forming the Linked Open Data (LOD) cloud. Most exist by means of subsidies from research projects or governments.[2] In contrast to the traditional Web, the traditional incentive mechanisms to publish data do not work in the WoD, as datasets are usually queried through algorithms rather the viewed by people. As a consequence, results of WoD queries often do not contain any attribution to the original source removing most non-monetary benefit to the data publisher. The lacking attribution nullifies many of the possible motivations of the original Web: there is no fame in providing a join-attribute, no promulgation of ones thought in helping filter some dataset, and little interpersonal communication when sending triples over the web. Indeed, even advertising—one of the most common financial incentives on the traditional Web—seems moot in an environment optimized to filtering unwanted information. Some might argue that a solution to this problem would be to concentrate all data in one big database akin to Google's indexing of the whole web. Such a centralized source would be able to extract monopolistic fees from consumers and, hence, pay for data maintenance and provision – not necessary a desirable but a viable solution. But as Van Alstyne *et al.* [20] argue, incentive misalignments would lead to enormous data quality problems and, hence, inefficiencies removing this approach from consideration. *So how can the WoD wean itself from purely relying on government subsidies?*

In this paper we propose *the use of a global price-market for data* to address the questions of economic viability of the WoD. Specifically, we believe that such a market could close the gap between applications that need access to diverse datasets and data providers whilst providing a well-defined set of incentive mechanisms to all involved parties. End-users would send queries to the marketplace platform that would negotiate with data providers for answers. As a side-effect the market-mechanism would efficiently allocate the data sources necessary and capable in answering a given query within the quality constraints set.

The contributions of this paper are the following: First, we believe that this is the first paper that models querying on the WoD as a market designed to set clear incentives for all participants. Second, *we simulate a market-based SPARQL broker* that receives query requests, budget allocations, and quality constraints from an application, runs an auction among possible providers, and then returns a result set within the budget and quality constraints defined. Third, we *show that such a market-place—if designed properly—will not only serve paying applications but also serve non-paying applications* (i.e., answering free queries). Indeed, our simulations show that *a properly designed market-place benefits from*

[1] http://richard.cyganiak.de/2007/10/lod/9 and http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[2] the uk open linked data initiative and others

the provision of some free data in terms of welfare for the overall society, consumer welfare for applications asking queries,[3] and to data providers. We even show that if designed properly a global price-market for data could incent commercial data providers to subsidize (or run) free data providers, as their existence will raise the commercial data providers profits.

2 Is Market-based Query-processing viable?

The main scientific contribution of this paper is to show that a market-based SPARQL broker can offer an efficient and incentive compatible (or ex-post individually rational) mechanism for allocating data providers on the WoD. In addition, we argued in contributions that such a mechanism—if appropriately designed—would entice all players (that is: queriers or consumers and data providers or producers) to embrace the availability of a mix of both commercial and free data providers. We operationalize this research question in four hypotheses groups. The first takes the perspective of the overall market; the second and third group establishes the individual rationality of the consumers and data providers respectively. The fourth and last set of hypotheses goes even further in arguing that commercial data providers could be enticed to pay for the free data provision by their increased profit resulting from their availability. In the following we elaborate on each of the hypothesis groups.

The coexistence of profit-oriented/commercial and free data providers is preferable to "society" over the existence of only commercial providers

The provision of data for free lies at the heart of the LOD movement. LOD proponents essentially argue that the benefits of publishing data for free induces network effects that by far offset the cost of the free provision [5]. Providing "only" free data, however, has the drawback that the societal welfare benefits from these data remain proportional to the initial subsidies. It does not provide the proper incentives to publish, additional, non-subsidized data. In addition, as Van Alstyne *et al.* [20] point out, free publication also raises the issue that publishers may not be interested in investing in quality maintenance. This is comparable to the *tragedy of the commons* effect, where nobody feels responsible to preserve and extend the existing data sources. Lastly, it is also unclear if a centralized subsidy mechanism is the right approach to determine which parts of a WoD require investments and extensions. Hence, we limit our analysis to a purely commercial setting and a mixed setup, where profit-oriented (or commercial) providers coexist with free, potentially-subsidized providers. To compare these two settings we hypothesize:

[3] Whilst it seems obvious that consumers profit from free data provision, an environment without any commercial data providers will only offer limited benefits to consumers, as data providers have no incentive to invest in their services beyond the subsidies. In the absence of subsidies data providers could even shut down their publication due to running costs.

Hypothesis 1a: Mixed data provision (costly and free) will lead to higher overall welfare, where we operationalize welfare as the sum of the profits of data providers and the consumer surplus as the budget allocated to answer queries minus the cost paid for the query results.

Hypothesis 1b: Mixed data provision will lead to higher overall quality in provided results

The coexistence of profit-oriented/commercial and free data providers is preferable to "consumers" over the existence of only commercial providers

When focusing on consumers, only the free access to all data seems like a good idea. As argued above, however, there are caveats. Assuming that access to data is limited by some kind of limited resource free access, will likely overburden that resource and the provision quality will deteriorate, just like the quality of many public goods (eg., streets) seem to suffer (eg., congestion or lack of servicing) unless some resource-considerate allocation mechanism (eg., road pricing) is installed. As we will see, our market mechanism will only find queries that fit the minimal quality defined by the consumer. Hence, we investigate the following hypothesis:

Hypothesis 2a: Mixed data provision (costly and free) given a predefined minimal quality will lead to higher consumer surplus, where we operationalize consumer surplus as the maximum budget the consumer is willing to pay to receive the answer minus the actual cost paid.

Also, we assume that in a universe of mixed free and commercial data providers, free data provision will not have the same reach in terms of the kind of data they offer. Also, some data may not be available at a desired quality level, as the free data providers offering them may drive the commercial providers out of business. Hence:

Hypothesis 2b: Mixed data provision leads to an overall lower recall for consumers, as some consumers will only query the free providers.

The coexistence of commercial and free data providers is preferable to commercial data providers over the existence of only commercial providers

Most people would intuitively argue that competition by free providers is bad for commercial providers. We argue that this depends on the market structure. Consider a market where the value of one good is dependent on the availability of another. As Parker and Van Alstyne [14] show, such a situation may entice companies to give away one of the goods in order to drive up the profit of another. A practical examples of such a situation is the free availability of Adobe Acrobat Reader. But does this situation also apply to the WoD? We believe it does. Just consider a commercial data provider who sells a data set linking two available free datasets. His data is considerably more valuable in the presence of the free data sources than in their absence. This leads to our third set of hypotheses:

Hypothesis 3a: Mixed data provision (costly and free) will lead to higher overall profit for commercial providers. Since this is a two-sided market, we need to

show that not only the requesters profit from the introduction of free providers but also the providers. Hence, we need to examine the total profit of all providers with and without free providers.
In a true market, however, not all providers are equal. Some provide higher quality goods, others lower quality ones. We, therefore, introduce two provider types offering high-quality or low-quality data. We assume that serving high-quality data causes high costs, whereas low-quality can be provided at low costs. How does the presence of free providers endanger the competitive situation of high-quality and/or low-quality providers? Does it affect them similarly or in different ways? We investigate these questions in the following two additional hypotheses:
Hypothesis 3b: Mixed data provision (costly and free) will lead to higher overall profit for high-quality (and therefore high-cost) commercial providers.
Hypothesis 3c: Mixed data provision (costly and free) will lead to higher overall profit for low-quality (and therefore low-cost) commercial providers.
Commercial data providers are likely to cross-subsidize free data.
The last hypothesis tests if the increase in profits made by commercial providers in the presence of free data providers may be sufficient incentive for them to actually provide free data themselves or subsidize other providers to do so.

Hypothesis 4: The additional profits gathered by commercial providers in the presence of free providers is greater than the cost of running the free providers.

3 Related Work

Efficient resource allocation is a major research topic of distributes systems. This section summarizes the most relevant work in *computational economics* or *agorics*, *databases* and the *semantic data management* research directions.
Economics-based computing. Early research on microeconomic-based scheduling focused on the efficiency of computational resources allocation. In Enterprise [13], e.g., tasks are efficiently allocated between LAN connected nodes. Employing a market metaphor, task processors broadcast *request for bids* and *bid* on tasks, where bids reflect task completion times. Likewise, Spawn [21] utilizes a market mechanism to optimize the use of idle computational resources in a distributed network of heterogeneous workstations. More recently, Tycoon [10], a distributed computation cluster, features a market-enhanced proportional-share resource allocation model. The authors claim that an economic mechanism is vital for large scale resource allocation – a common problem on the Web. Furthermore, market-based optimizations have proved to be as good or better than traditional allocation methods in grid-computing schedulers with the added benefit of pluggable pricing models, objective functions, and access policies. Similarly, Auyoung et al. [1] demonstrate how profit-aware algorithms outperform non-profit aware schedulers across a broad range of scenarios.

Microeconomic-based optimizations have been shown to be successful in areas such as discussion-forum optimization [12], P2P systems [15], and real-time query answering systems [9], where they greatly improved performance, when demand reaches its peak. Stemming from the interaction between data management and

microeconomics, Mariposa [18]—a WAN-scale RDBMS—drops the traditional cost-based optimizer in favor of a market-based one. Since common assumptions like: static data allocation, single administrative structure, uniformity of network and site capabilities, do not hold at WAN scale, Mariposa binds data objects to owners and assigns budgets to queries. As each site tries to maximize its revenue, results are the outcome of broker-mediated auctions.

Van Alstyne et al. [20] focus on the impact of *soft*, intangible factors such as *ownership* as key mechanisms for incentive-provisioning in database systems. They find that in absence of explicit contracts, which reward those that create and maintain data, ownership is the best way to incentivize data creation and maintenance. Consequently, while technical hurdles for database integration can be overcome successfully, ignoring appropriate incentive structures can just as well lead to system failure.

Distributed Data Management on the Web of Data. A wide spectrum of approaches for SPARQL processing on the WoD exist. Early methods borrowed from traditional distributed query processing [19, 2], while others build on top of P2P systems[4]. Although performant, these approaches lack with respect to their *heterogeneity and openness* characteristic that is so prevalent in the WoD. One of the earliest RDF federations, DARQ [16], makes use of a less restrictive query decomposition model mandated by its predicate data partitioning model. Similarly, SemWIQ [11] relies on concept-based data partitioning, where sources are selected based on the concept's `rdf:type`. More recently, several Sesame extensions provide federated SPARQL query answering capabilities such as AliBaba[4] and FedX [17]. The latter improves query execution by focusing optimizations on the join operators. Relying on precomputed statistics exposed as *VoiD*[5] descriptors, SPLENDID [6] strives to achieve the same goal. In contrast to these approaches, Avalanche [3] strives to achieve this target by combining query-time discovery with a competitive and parallel multi-plan execution strategy. It does, however, not guarantee the provision of complete results.

Discovering resources at runtime Hartig et al. [7] describe an approach for executing SPARQL queries over LoD based on link traversal. Whilst the technique embraces the WoD's flexibility to its full extent a number of limitations still exist such as the impossibility to execute certain kinds of queries in conjunction with a significant drop in performance.

4 System Design

In this section we describe the overall system architecture and propose a microeconomics based extension of the adopted federated SPARQL execution pipeline.

Background As highlighted in Section 3 several federated SPARQL engines have been proposed to date. Due to the multi-tenant aspect of the WoD we did not consider "closed" federated RDF stores, where a centralized agent is selecting the

[4] http://www.openrdf.org/alibaba.jsp
[5] http://www.w3.org/TR/void/

presumably best plan. We did strive for an approach, where multiple alternative plans could compete in the economic market-place. So we searched for a system that could devise a multitude of promising plans based purely on VoID-like statistics.

We found a solution in Avalanche [3], which uses VoID-like statistics to explore the space of all possible plans (or query decompositions) ordered by a likelihood of success. Hence, we extend Avalanche with a market-based plan selection. The resulting `MaTriX` system starts an auction for all plan fragments of the Avalanche plan universe to find plans that are not only feasible but can be executed within budget under a given quality constraint. Specifically, for a given Query A, Avalanche generates a universe U_A of $|U_A|$ plans ordered by their "promise" to generate results quickly. For each of the top K plans $P_i \in U_A$ meeting the minimum quality constraints, `MaTriX` identifies the constituting query fragments F_{P_i} and starts a reverse auction among providers for its provision. The query fragments of the winning plan according to the budget and quality constraints are then executed and composed to derive the answer.

Challenges and Requirements Given the messy nature and size of the Web of Data, `MaTriX` is faced with a number of challenges:

i. **Scale**: Since data-providers potentially bid on high numbers of query fragments, the platform must be able to scale to the volume of ongoing transactions.

ii. **Auctioning**: In common (forward) auctions, bids represent the valuation a bidder has for a certain good. In `MaTriX` however, bidding is occurring at the supply side as providers bid to cover their costs. Hence, we run a so-called *reverse* or *procurement auction*, where the providers get at least their bid + Δ (or profit)[6].

iii. **Anonymity or Auction Complexity**: In most settings, auction participants act independently and compete against all other bidders. Depending on the auction type, participants may be able to learn how to adapt the bid strategically to win subsequent auctions. In order to ensure an efficient marketplace `MaTriX` needs to prevent strategic bidding by the use of an appropriate mechanism. Currently, bidders are part of anonymous groups. Hence, the losing bidders cannot strategize as easily, since it is unclear if the bidder itself or another (unknown) actor in the group was responsible for the failure.

Theoretical Considerations in the Light of the Challenges To understand how to fulfill the above challenges using an auction we require a set of concepts from auction theory. These are:

Forward First and Second Price Auction. In a forward auction, customers have a certain valuation for a good, and bid to be able to buy this good. In a First Price Auction (FPA) the customer with the highest bid wins and has to pay his bid. Since high bids lower a customer's profit it is rational for him to strategize and lower the bid. But lowering the bid increases the chances of losing the auction. Therefore, bidders are best off to bid $B(v) = \frac{N-1}{N} \times v$, where v is their true

[6] Δ can also be 0

valuation and N is the total number of bidders in the auction. In a second price auction (SPA) a customer with the highest bid wins but has to pay the bid of the second highest bidder. This has been shown to completely discourage strategic behavior and will induce bidding and the truthful valuation. Therefore, bidders are best off bidding $B(v) = v$. [8]
Reverse Auction. In a reverse auction, the role of the customer and seller are reversed. Customers have a maximum budget they are willing to spend for a certain good. Sellers have certain costs to provide this good and bid depending on their cost to sell this good. Again, auctions can be governed by FPA or SPA scheme, where the former leads to higher bids than their true costs and the latter induces truthful bidding by providers.
Reverse Auction with Bidding in an Anonymous Group. In our situation, we do not have a reverse auction for one indivisible good, but a auctions for different query fragment compositions. Here, the query with the lowest total costs wins the auction. Thus, it is not single bids that win the auction but the combined (or collective) bid of the whole group of providers assigned to the same query. To avoid collusion among providers they are neither supplied with any information about each other nor the overall plan they bid for. Nonetheless, in a reverse FPA with bidding in an anonymous group providers will still bid a higher amount than their actual costs to make a profit. In a reverse SPA this kind of behavior gets discouraged, as the payout depends on the second lowest bid. Unfortunately, the truthfulness is lost in this kind of SPA, since providers can increase their share by not truthfully reporting their costs. But since (i) we do not give any information about the size of the bidding group, the bids of the other participating providers or the total costs of the plan and (ii) all those factors change after each auction is should be difficult for providers to analyze if misreported cost improves or worsens their profit. Hence, we assume that their dominating strategy is to report their true costs.

Putting it all together – A Process Example Prior to any query execution, providers need to register with `MaTriX` by providing typical SPARQL endpoint information such as the address and simple statistics like vocabularies used. `MaTriX` then needs to assess the quality of a provider – a procedure beyond the scope of this paper (that could also be outsourced to a third-party quality assessment service). During an actual query execution `MaTriX` processes a query A with the budget B and minimal quality constraint Q in 3 phases (see Figure 1):

During the *planing and source selection phase* Avalanche generates the ordered list of plans U_A for query A. Then, during the *bidding and plan selection* phase `MaTriX` starts a reverse auction for each of the top k plans $P_i \in U_A$ by issuing requests for bids to the providers who have information pertinent to any query fragment F_{P_i} belonging to P_i. From all succeeded auctions `MaTriX` picks the plan P_i that minimizes the cost and is still under budget whilst providing a sufficient quality (i.e., $i : min_i[\sum(cost(F_{P_i})] < B \wedge min(quality(F_{P_i})) < Q$). Finally, `MaTriX` passes the winning plan P_i to Avalanche for execution and pay out the fees to the providers.

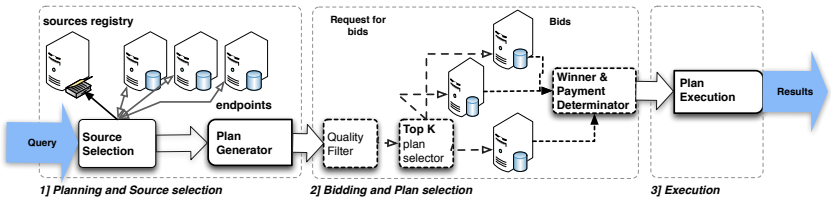

Fig. 1. General MaTriX architecture

5 Experimental Evaluation

In order to determine the economic viability of `MaTriX` , we designed a controlled simulation where different factors where observed in isolation as detailed next.

Simulation Setup Due to the lack of appropriate benchmarks appropriate for our question and the high effort/cost to create one our evaluation used simulated data providers. We based the simulation on the following assumptions: *i*) each provider is able to answer the SPARQL fragment assigned to it, *ii*) the cost per query to providers is fixed (arguably, this cost should be dependent on the query complexity and current load; but a fixed cost could make sense in a cloud-hosted environment, where one pays per CPU/IO usage) *iii*) queries are generalized and, likewise, simulated. For practical reasons, we set k to 10 when selecting the *top* k most "productive" plans and all queries are decomposed into 4 fragments.

Table 1. Experiment settings by auction type and provider distribution

Setting	Auction type	#HH	#LL	#F	Setting	Auction type	#HH	#LL	#F
FPA-8P	First Price	8	16	0	*FPA-8P-F*	First Price	8	8	8
SPA-8P	Second Price	8	16	0	*SPA-8P-F*	Second Price	8	8	8
FPA-4P	First Price	4	20	0	*FPA-4P-F*	First Price	4	8	12
SPA-4P	Second Price	4	20	0	*SPA-4P-F*	Second Price	4	8	12

Providers represent the supply side of the market and are reachable SPARQL endpoints. A total of 24 providers were simulated. Providers are classified into **low quality** (LL) when $Q \in [0, 5]$ and **high quality** (HH) if $Q \in [6, 10]$. We assume a strong positive correlation between the offered quality and costs: for low quality providers the cost $C \in [1, 50]$, for high quality $C \in [51, 100]$.[7] In addition, we also have free providers (F) that are assumed to provide low quality data. Whilst it is true that most free providers would rely on subsidies (and may offer high quality data) assuming low quality is a prudent assumption. To test our hypotheses we varied the distribution of providers. As detailed in Table 1 we either had 4 or 8 high-quality providers (designated by *4P* or *8P* in the setting name). In the settings with free providers (designated by *F*) we had 8 low-quality providers and in the others all the remaining providers where low-quality.

[7] Costs, bids, and budgets are represented in a virtual currency not mentioned in text for brevity. Quality is also assumed to be assessed on a virtual, ordinal scale.

Provider Bidding Strategies. In a FPA auction providers will bid strategically by starting above their costs and adapting on the feedback they get from `MaTriX`: if they lose, then next bid is lowered; else it is incremented. Note that providers get feedback only after an auction finishes, even if they bid several times. In addition, due to the out of order completion of concurrent auctions the same effect can appear. Hence, adaptation only happens between queries. For SPA auctions we assume that bidders will bid their true costs and not act strategically—a common outcome in second price auctions.
Requesters represents the demand side of the market. They are the customers that submit SPARQL queries. We assume that requesters manifest certain preferences regarding the quality of requested results, where the requested quality Q is supposed to be less or equal to the returned result quality Q_R. Additionally, their budget is either **low** where $B \in [200, 250]$ or **high** for $B \in [250, 400]$. To model an environment as diverse as possible, we create four requester types populating the two dimensional binary space of quality Q and budget B (*high-high, high-low, low-low, low-high*).
Requests are issued by the submission of the tuple $<$ SPARQL $Query, Q, B >$, where Q and B are quality and budgetary constraints.
Experiments. For each setting described in Table 1 we first initialized the correct number of provider types by randomly drawing their cost and quality from a normal distribution in the ranges mentioned above. Second, we generated 1000 requests that where randomly assigned to one of the four types of requesters. Again, we drew the budget and quality constraints from a normal distributions in the appropriate ranges mentioned above. Third, we ran the sequence of 1000 request through `MaTriX` and observed the bids, revenue, and costs for providers and, requests, price, and quality received form the requestors. To ensure that we did not get statistical outliers we ran this procedure 50 times for each setting.

Empirical hypotheses testing and validation

Figure 2 shows the summary results for testing Hypotheses 1a and 1b. The two graphs on the left clearly show an increased welfare (as the sum of the profits of data providers and the consumer surplus as the budget allocated to answer queries minus the cost paid for the query results) when comparing the setting with free providers (red) to the setting without free providers (blue). The difference is significant as confirmed by a Mann-Whitney U test ($p \in \{$9.70e-16, 2.01e-16, 6.25e-15, 7.15e-17$\}$). We can, hence confirm Hypothesis 1a.
As the righthand side graphs show, the introduction of free providers does not perceptibly change the answer quality (Mann-Whitney U $p \in \{$7.66e-01, 8.35e-01, 5.33e-01, 2.18e-01$\}$). Whilst this rejects Hypothesis 1b it's actually a good result, as it holds that free providers will not lead to a significant change in answer quality.

Figure 3 shows the summary results for testing Hypotheses 2a and 2b. The two left graphs show a significantly increased consumer surplus (as the difference between allocated budget and paid price; MWU: $p \in \{$3.98e-18, 3.53e-18, 1.60e-16, 3.53e-18$\}$) when comparing the setting with free providers (red) compared with the setting without free providers (blue). Note that the effect is much higher

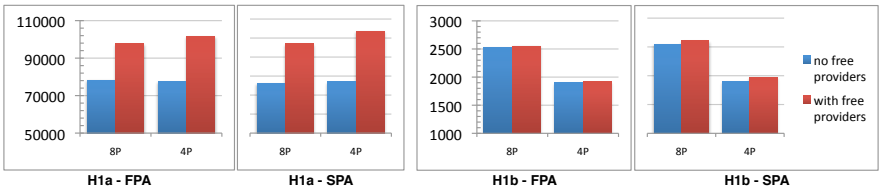

Fig. 2. Hypotheses 1: Total Welfare and Answer Quality

in the SPA (40% increase) compared to the FPA setting (37% increase) for the 4P case, while in the 8P case the situation is reversed (FPA with 32% vs. SPA with 28% increase). Therefore by introducing more free providers the consumer surplus increases more in the SPA setting, where providers are assumed to bid their truthful valuation. On the right the figure graphs the recall, or the number of queries answered in each of the settings. We can see that the introduction of free providers does not perceptibly change the recall (Mann-Whitney U $p \in \{3.35\text{e-}01, 3.26\text{e-}01, 5.26\text{e-}02, 1.60\text{e-}01\}$). Again, whilst this rejects Hypothesis 2b it is actually a good result, as it maintains that free providers will not lead to a significant change in number of queries answered (e.g., due to budget or quality constraints).

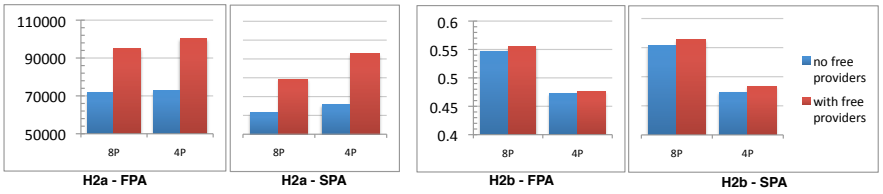

Fig. 3. Hypotheses 2: Total Consumer Surplus and Recall (fraction of answered queries)

Figure 4 shows the summary results for testing Hypotheses 3a-c. As we can see the situation here is somewhat more complicated. Indeed, in the FPA settings the overall profit actually shrinks with the exception. In the SPA setting the profit either increases significantly (MWU: $p = 1.47\text{e-}13$ for 8P) or is statistically, indistinguishable (MWU: $p = 6.96\text{e-}02$ for 4P). We, hence, have to reject or accept Hypothesis 3a depending on the setting. Note that this result is even more surprising as they are less overall commercial suppliers in the settings with free providers.

As the Figure shows on the right the inclusion of free providers raises the profit of high-quality commercial providers in the SPA case (MWU: $p \in \{3.53\text{e-}18, 3.53\text{e-}18\}$). In the FPA case, in contrast, the profit increase is insignificant (MWU: $p \in \{ 4.99\text{e-}01, 1.18\text{e-}01\}$). Hence, they will be strong supporters of a `MaTriX`-like setup operating with a SPA with free providers. The low-quality commercial providers, however, will suffer from the increased competition of the free low quality free providers (bottom two graphs of Figure), gain a lower profit with their introduction and, hence, oppose an introduction of both subsidized free providers and a `MaTriX`-like system.

As a consequence, we can infer that the introduction of free providers is highly beneficial to high-quality commercial providers in the presence of a `MaTriX`-like setup operating with a SPA. They may even be prepared to cross-subsidize free providers in order to reap the higher profits—a question we address in the evaluation of Hypothesis 4 below. For low quality providers the situation is less advantageous. They loose profits from the introduction but will still remain in the market as they can still garner profits. Actually, arguing further, it may provide the proper incentive to low quality providers to improve their data quality in order to become high quality (and thus high-profit) providers—a societally desirable result.

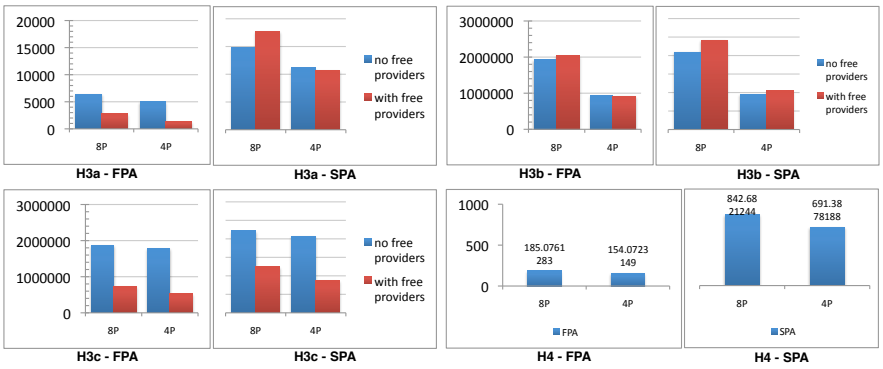

Fig. 4. Hypotheses 3: Profits normalized to one provider

To investigate Hypothesis 4 we computed the maximum amount the high-quality commercial providers would be prepared to pay free providers for answering a query as follows: for each experiment we took the total profit gained by the high-quality providers in settings with free providers and subtracted from it the profit in the analogous setting without free providers. The resulting difference (the total additional profit in markets with free providers) was then divided by the number of queries served by free providers resulting in the maximum subsidy the high-quality commercial providers would be prepared to pay per free query. As the results graphed in the lower right group of Figure 4 show, these numbers are well above the maximum cost of 100 incurred by any provider when servicing one query in the FPA and several times over in the SPA settings. We can, hence, confirm Hypothesis 4 that the *high-quality providers should be prepared to pay for the cost of queries served by the free data providers as it increases their own profit*. This is especially true in the SPA setting, where their profit after covering the cost of free queries is a multiple of those subsidies.

6 Limitations and Future Work

`MaTriX` is a first attempt to investigate the feasibility of market-structures for financing the WoD. The conclusions we can draw from these exploration are

limited by the generalizability of both the underlying `MaTriX` assumption and its evaluation.

Limitations of the `MaTriX` Setup: The current `MaTriX` setup assumes that *the market-platform itself is provided for free*. We did not incorporate any structure that would tax market participants to raise any funds for its running but do not expect it to deviate from existing maker-designs where a market-provider collects some fee.
Also, the current `MaTriX` market design tries to limit the need for elaborate reasoning and strategizing on the side of providers by (i) supplying only limited information about the queries and (2) clarifying the payout rules. We did not, however, show that such strategic behavior would not be beneficial to the providers and can, hence, not rely on the providers being truthful. Such behavior has been shown to be dominants in a traditional SPA. But in our complex reverse situation such a proof is still outstanding and maybe even impossible – a task we will turn our attention to in the near future.
The current version of `MaTriX` also relies on Avalanche as its plan generator. Since Avalanche is not omniscient it relies on the availability of VoiD-like statistics about participating data-sources. Whilst assuming the existence of such statistics is reasonable the plans Avalanche devises could still have en empty result set. Otherwise, Avalanche would be capable of running `ASK` queries without accessing the data. But how will consumers react when getting an empty dataset for a query they paid for? In particular, given the open world nature of the WoD such a result could not even be taken as a negative answer. But does such an empty answer still contain sufficient information to be of value? These question will need to be addressed in future investigations.
Another limitation of the reliance on Avalanche is that the providers do not self-select into query fragments they want to participate in. In essence, `MaTriX` could send out the whole query to all providers who have some information about some part of the query and the providers could query for any combination of fragments. Whilst this approach would simplify `MaTriX`'s task at a first glance it would significantly complicate the provider's bidding task and the aggregation of the processing of the bids, as `MaTriX` would have to evaluate a combinatorial space of query fragment bids. Earlier versions of Avalanche employed this approach and where hampered with serious scalability issues. Nonetheless, we need investigate if this approach would be desirable under certain circumstances.
Finally, the current model assumes a centralized quality assessment of the query results based on the information/answers supplied by the providers. As in all centralized approaches the centralized assessment is simpler and more difficult to manipulate but may not scale sufficiently. We will, therefore, have to investigate distributed quality assessment approaches akin to the methods used in online marketplaces such as eBay.

Limitations of our evaluation: The main limitation of our current simulation lies in the limited capabilities of the providers. In the real world, we would not expect providers to self-select into high- and low-price groups but to test the market and eventually shift to the price point optimal for their data offering.

Also, our simulation assumes that providers have a fixed cost in answering queries. This assumption is somewhat problematic: in the real world server-cost is more alike to a step function (buying an additional server is costly; answering an additional query on a server with free capacity is cheap) and depends on the complexity of a query. Hence, we will either need to extend our simulation to include a more elaborate (potentially learned) cost function for providers that will influence their bidding strategy or run the evaluation with real servers that will influence their bidding based on their past and current load.
Furthermore, we assume that the provider cost and their provided quality correlate. It is unclear if this assumption is correct, as even low-quality information may be costly to come by and still valuable. Further investigation will have to show how sensitive our findings are to this assumption.
Finally, our current simulation keeps the providers in the dark about the number of fragments that are combined into a query request. Initially, we had thought that providers should be kept in the dark in order to elicit truthful behavior. If providers were, however, trying to strategize they may choose to lower the price for queries with many fragments in order to increase the probability of gaining the "contract" to provide the service and vice-versa. Again, as mentioned already, the whole issue of provider strategic behavior needs to be investigated holistically in a future study.

7 Conclusions

In order to become sustainable the WoD needs to find means for financing itself. Monetization approaches from the traditional web do not seem to be suitable for it, as they rely on people consuming the information fragments (with associated advertising) rather then only a complied summary. To address this shortcoming we propose the use of a market mechanism called `MaTriX` for the WoD to wean itself from a subsidy-oriented financial foundation.

We show using a simulation that `MaTriX`'s reverse, sealed-bid second price auction mechanism provides consumers with a higher consumer surplus, in a mixed profit-oriented and free provider setting. Whilst this may not be surprising, we also show that producers will be able to reap higher profits in the mixed setting. We even find, that for profit producers might be enticed by these higher profits to cross-subsidize the free information providers – a surprising result. Furthermore, another positive aspect is denoted by the general incentive for providers to expose high(er) quality data which inturn drives profits up.

Whilst our findings are clearly preliminary and burdened by a number of limitations our papers presents the first systematic study trying to provide a sound economic foundation for the WoD. Indeed, the study lays out the foundation and agenda for such economic studies of data on the web. As such it paves the way to a financially healthy WoD.

References

1. L. G. Alvin Auyoung. Service contracts and aggregate utility functions.

2. J. U. Andreas Harth. Yars2: A federated repository for querying graph structured data from the web. 2007.
3. C. Basca and A. Bernstein. Avalanche: Putting the Spirit of the Web back into Semantic Web Querying. In *The 6th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2010)*, Nov. 2010.
4. M. Cai and M. R. Frank. Rdfpeers: a scalable distributed RDF repository based on a structured peer-to-peer network. In *13th International World Wide Web Conference (WWW)*, pages 650–657, 2004.
5. A. Golliez, C. Aschwanden, C. Bretscher, A. Bernstein, P. Farago, S. Krügel, F. Frei, B. Bucher, A. Neuroni, and R. Riedl. *Open Government Data Studie Schweiz*. Berner Fachhochschule, Bern, 2012.
6. O. Görlitz and S. Staab. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. In *Proceedings of the 2nd International Workshop on Consuming Linked Data*, 2011.
7. O. Hartig, C. Bizer, and J.-C. Freytag. Executing SPARQL Queries over the Web of Linked Data. *8th International Semantic Web Conference ISWC2009*, pages 293–309, 2009.
8. V. Krishna. *Auction Theory*. Academic Press, 2010.
9. A. Labrinidis, H. Qu, and J. Xu. Quality contracts for real-time enterprises. pages 143–156, Sept. 2006.
10. K. Lai, L. Rasmusson, E. Adar, L. Zhang, and B. A. Huberman. Tycoon: An implementation of a distributed, market-based resource allocation system. *Multiagent and Grid Systems*, 1(3):169–182, Aug. 2005.
11. A. Langegger, W. Wöß, and M. Blöchl. A semantic web middleware for virtual data integration on the web. *Lecture Notes In Computer Science*, 2008.
12. N. López, M. Núñez, I. Rodríguez, and F. Rubio. Encouraging knowledge exchange in discussion forums by market-oriented mechanisms. In *Proceedings of the 2004 ACM symposium on Applied computing - SAC '04*, page 952, New York, New York, USA, Mar. 2004. ACM Press.
13. T. W. Malone, R. E. Fikes, and M. T. Howard. Enterprise : a market-like task scheduler for distributed computing environments, Nov. 1983.
14. G. G. Parker and M. W. van Alstyne. Two-sided network effects: A theory of information product design. *Management Science*, 51(10):1494–1504, 2005.
15. J.-A. Quiané-Ruiz, P. Lamarre, S. Cazalens, and P. Valduriez. Managing virtual money for satisfaction and scale up in P2P systems. In *Proceedings of the 2008 international workshop on Data management in peer-to-peer systems - DaMaP '08*, pages 67–74, New York, New York, USA, Mar. 2008. ACM Press.
16. B. Quilitz and U. Leser. Querying Distributed RDF Data Sources with SPARQL.
17. A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. FedX: optimization techniques for federated query processing on linked data. pages 601–616, Oct. 2011.
18. M. Stonebraker, P. M. Aoki, W. Litwin, A. Pfeffer, A. Sah, J. Sidell, C. Staelin, and A. Yu. Mariposa: a wide-area distributed database system. *The VLDB Journal The International Journal on Very Large Data Bases*, 5(1):48–63, Jan. 1996.
19. H. Stuckenschmidt, R. Vdovjak, G. J. Houben, and J. Broekstra. Index structures and algorithms for querying distributed RDF repositories. In *13th International World Wide Web Conference (WWW)*, May 2004.
20. M. Van Alstyne, E. Brynjolfsson, and S. Madnick. Why not one big database? principles for data ownership. *Decis. Support Syst.*, 15(4):267–284, Dec. 1995.
21. C. Waldspurger, T. Hogg, B. Huberman, J. Kephart, and W. Stornetta. Spawn: a distributed computational economy. *IEEE Transactions on Software Engineering*, 18(2):103–117, 1992.

M. Zollinger et al.: Market-based SPARQL brokerage with MaTriX: towards a mechanism for eco
Technical Report No. IFI-2013.04, June 2013
Dynamic and Distributed Information Systems (DDIS)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
URL: http://www.ifi.uzh.ch/