Management of collaborations in digital marketplaces

Lu Zhang
Systems and Networking (SNE) lab
University of Amsterdam
Amsterdam, Netherlands
l.zhang2@uva.nl

Dissertation Advisor(s): Prof. Dr. Cees de Laat, Dr. Paola Grosso
Dissertation Committee Members: Not yet formed

DOCTORAL DISSERTATION COLLOQUIUM

EXTENDED ABSTRACT

***Abstract*— With everyone generating value out of data, our work focuses to distributed data trading platforms, Digital Market Places (DMPs), that can handle the intricacies of data sharing, e.g. how, where, and what can be done with the traded data. Here we represent collaborations among involving parities in DMPs in the form of archetypes and model them with numeric representations for easier manipulation with standard mathematical tools. We also develop a methodology which aims to select a best-fit infrastructure archetype with any customer-defined application request. In addition, we propose multiple metrics which allows evaluate and compare competing DMPs systemically from more dimensions: coverage, extensibility, precision and flexibility.**

Keywords-component; Open Data Market, Data Marketplace, Trusted Data Market, Industrial Data Space, Data Economics, STREAM Data Properties

I. INTRODUCTION

In the era of big data, the amount of collected data is increasing dramatically [1]. Sharing and utilizing such data can generate great value and improve collaborations among parties. But security and privacy concerns may rise, especially in scenarios that members are normally competing with each other.

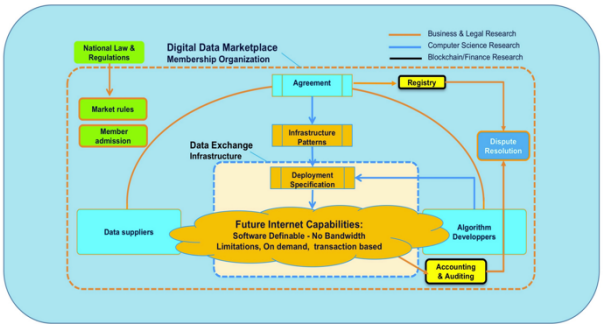

Figure 1 A high level framework of a DMP

Newly emerging Digital Marketplaces (DMP) aims to facilitate such trusted data sharing for a specific purpose [2-4]. A DMP is a membership organization to support members to achieve a common goal by data asset sharing. Figure 1 illustrates a high level framework of a DMP. The movements and execution of data and compute are governed by an *Agreement* achieved by all members in this DMP instance. The *Infrastructure Pattern* is dependent on concrete *Agreement* for each DMP instance and those rules are enforced by underlying *Data Exchange Infrastructure* with future network capabilities. Here begs a question:

How to create such a platform for sharing data and compute governed by the agreed policy?

We try to answer, at least part of, the question for my PhD dissertation. Some work has been done and more needs to be investigated in the future of my PhD life.

II. MODELING OF MULTI-PARTY COLLABORATIONS

Collaboration models are defined to describe restrictions about how the data is accessed, shared and used during a collaboration, which are included in *Agreement.* They serve a role in connecting policies to the underlying digital infrastructure. Normally, collaboration models are defined from both DMP operator side and a potential customer side. We call them *Archetypes* and *Application Requests* respectively. An *application request* may comprise both hard requests and soft requests. Hard requests are not negotiable and must be fulfilled in the collaboration process. However soft requests could be adjusted to better fit any existing collaboration archetype.

In order to manage these multi-party collaborations, we should first model them properly. A multi-party collaboration relationship can be fully described by four attributes: *Source, Target, Collaboration level, Collaboration scope. Collaboration level* represents the manner of collaborating

among members and *Collaboration scope* describes which resource could be shared between specific parties [5]. Parties in the DMP may collaborate across multiple scopes, data, algorithm and intermediate result.

As illustrated in Figure 2, a multi-party collaboration relationship is effectively modelled as a 3D matrix. Under each scope along z-axis, the collaborations are represented as a 2D matrix with corresponding *collaboration levels*. With such numeric representations, we can easily manipulate collaborations with standard mathematical tools.

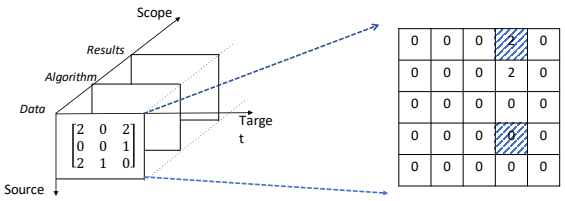

Figure 2: Modeling of a multi-party collaboration relationship. On the left we see the relations between sources and targets for the three scopes; on the right we zoom in on one specific scope, where the crossed-out cells represent hard requests.

III. A SELECTION ALGORITHM OF DMP ARCHETYPES

A DMP normally supports multiple archetypes to allow customers to choose from. Also, a potential DMP customer may have different collaboration requests for different applications. So he/she has to participate different DMP instances to meet the requirements. It is highly beneficial to develop an algorithm for selecting best-fit archetype automatically.

We define similarity measures between collaboration models, which is effectively quantified as a distance metric. Either an archetype or an application request can be mapped as a point in a discrete space by calculating their mutual distances.

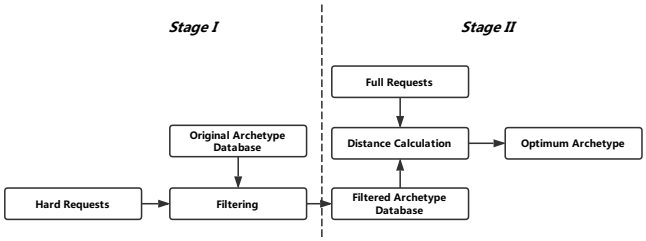

Figure 3: Flow chart of the archetype selection algorithm. Stage I is concerned with filtering the archetypes based on hard requests, and Stage II calculating distances to identify the optimal archetype

The matching algorithm consists of two stages, filtering (Stage I) and archetype selection (Stage II). Figure 3 describes the algorithm flowchart. At Stage I, all collaboration archetypes from *Original Archetype Database* are filtered with *Hard Requests* given by a potential customer. After *Filtering,* a subset of archetypes is kept in *Filtered Archetype Database* for further processing and the corresponding searching space shrinks. All the remaining archetypes are acceptable by potential customers for the compliance with *Hard Requests.* At Stage II, we first calculate the distances between *Full Application Request* and remaining archetypes in *Filtered Archetype Database*. Then select the *optimal archetype* as the one with minimum distance towards *Full Application Request*.

The definition of distance between archetype and application request is based on Weighted Hamming Distance and both collaboration models are pre-processed for more commensurable comparison [6].

Hence it is possible to identify the closeness between any application request to archetypes and match an application to a "closest" archetype.

IV. DMP EVALUATION METRICS

For potential customers it is interesting to know a-priori how easily one of their application requests can be fulfilled by a particular DMP; for DMP operators it is important to assess how well they can serve their user base generally. Here we propose multiple metrics that allow more nearly complete evaluation of a DMP:

- *Coverage*: How well the overall application requests can be satisfied by a DMP with a certain mismatch.
- *DMP Extensibility*: What is the potential richness of a DMP by decomposing and composing collaboration archetypes.
- *Application Extensibility*: How elastic of an application request in achieving a perfect match towards a given DMP.
- *Precision*: How well the supported collaboration archetypes of a DMP fit an application request.
- *Flexibility*: How easily an application request can be satisfied generally.

Metrics like *Coverage* and *DMP extensibility* are not related to individual request but represent a general feature of a DMP. However, *precision*, *flexibility* and *application extensibility* depend on both defined application requests and DMP itself.

Besides conceptual definitions, we also define concrete quantization methods for each metric. Here we take *coverage* as an example.

According to definition, *coverage* is highly dependent on how we define customer satisfaction. In our work, a potential customer is considered as satisfied if the distance, between his/her application request and optimum archetype in the DMP, is not larger than a pre-defined value. We call the parameter affordable distance D_A.

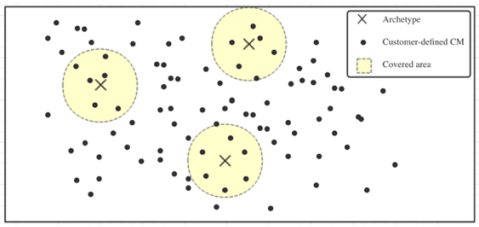

Figure 4: Illustration of coverage in a discrete space, with archetypes identified as crosses, application requests as dots, and covered areas represented by the yellow circles.

As illustrated in Figure 4, the covered area of an archetype is modeled as a sphere with radius of the D_A. Total covered area of multiple collaboration archetypes is the union of individual covered area. The metric *coverage* is quantified as percentage of the application requests, that fall into the covered area of supported archetypes, over the total number of collaboration models.

V. CURRENT RESULTS

We validate the effectiveness of all the metrics with DL4LD to show how these methodologies are applied to in real world [7]. To be consistent with Section IV we also discuss *coverage* as an example. Figure 5 describes the *coverage* with D_A as 4 and 6 respectively. Each group represents *coverages* of DMPs supporting archetype sets with equal size. It is shown that *coverage* increases approximately in a linear manner with larger archetype set size.

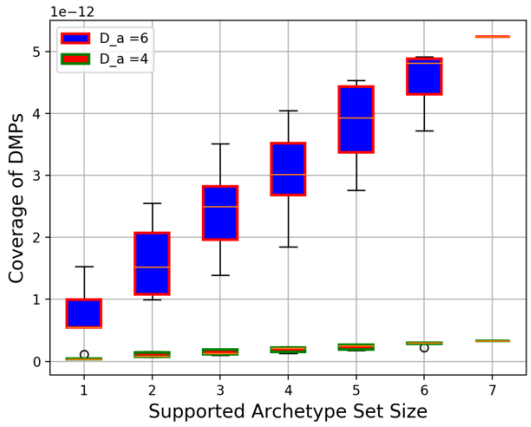

Figure 5: Coverage as function of increasing archetype set size in the DL4LD project for various affordable distance D_A

In addition, by analyzing values of those metrics, a DMP operator may find more optimized solution between implementation cost and achieved *coverage*. Shown in Figure 5, most inter-quartile range boxes have overlap values with their neighbors. This indicates that a DMP, who supports a larger number of archetypes, may result in a relatively lower *coverage*. One DMP operator may benefit from selecting a specific archetype set who has higher *coverage* but lower archetype size.

VI. CONCLUSIONS AND FUTURE WORK

We presented a model for describing DMP collaborations between participating parties. We showed that if the DMP collaboration archetype and the application request are consistently described we can map them together. This mapping allows us to identify the closeness of request and the offered infrastructure. We showed that the evaluation and comparison of competing DMPs are allowed and supported by having consistent and generic metrics, namely *coverage, extensibility, precision and flexibility.*

There are many interesting directions to investigate in the future. Firstly, the archetype selection procedure can be further improved. We can also integrate security and performance considerations to facilitate a multi-criteria decision making.
It is also interesting to develop a generic methodology to analyze risks with a given scenario, minimize those risks by applying state-of-art defense mechanisms and estimate the achievable security level quantitatively or qualitatively.

BIOGRAPHY

Dr. Paola Grosso is associate professor in Systems and Networking Lab (SNE) at UvA. Her research interests lie in the creation of sustainable e-Infrastructures, relying on the provisioning and design of programmable networks. She currently participates in several national projects, such as SARNET, DL4LD and EPI. See: https://staff.fnwi.uva.nl/p.grosso/

Prof. de Laat chairs the System and Network Engineering (SNE) laboratory at UvA. His group is/was part of a.o. EU projects GN4-2, SWITCH, CYCLONE, ENVRIplus and ENVRI, Geysers, NOVI, NEXTGRID, EGEE, and national projects DL4LD, SARNET, COMMIT, GIGAport and VL-e. See:http://delaat.net/

Lu Zhang is currently a PhD student in Systems and Networking Lab (SNE) at University of Amsterdam. She received her B.SC. and M.Sc from Shandong University, China and RWTH Aachen University, Germany. Her research interests include information security, container networks and novel networking infrastructures.

REFERENCES

[1] S. Sagiroglu and D. Sinanc, "Big data: A review," in 2013 International Conference on Collaboration Technologies and Systems (CTS). IEEE, 2013, pp. 42–47.

[2] S.Liebowitz,"Rethinking the networked economy:The true forces driving the digital marketplace," AMACOM Div. American Management Association, Dallas, 2002.

[3] A. Zerdick, K. Schrape, A. Artope, K. Goldhammer, U. T. Lange, E. Vierkant, E. Lopez-Escobar, and R. Silverstone, E-conomics: Strategies for the Digital Marketplace. Springer Science & Business Media, 2013.

[4] S.Cisneros-Cabrera, A.Ramzan, P.Sampaio, and N.Mehandjiev, "Digital marketplaces for industry 4.0: a survey and gap analysis," in Working Conference on Virtual Enterprises. Springer, 2017, pp. 18–27.

[5] A. Jøsang, E. Gray, and M. Kinateder, "Simplification and analysis of transitive trust networks," Web Intelligence and Agent Systems: An International Journal, vol. 4, no. 2, pp. 139–161, 2006.

[6] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov, "Hamming distance metric learning," in Advances in neural information processing systems, 2012, pp. 1061–1069.

[7] T. D. consortium. (2018) Data logistics for logistics data. [Online]. Available: https://www.dl4ld.net.