"Storytelling" in the economic LOD: the case of publicspending.gr

M. Vafopoulos, I. Anagnostopoulos, M. Meimaris, M. Klonaras, I. Xidias, A. Papantoniou, G. Vafeiadis, V. Loumos and G. Alexiou

Multimedia Technology Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens

Keywords: *open data, e-government, public expenditure, Ontologies, citizen empowerment, data journalism*

Abstract: The scope of the publicspending.gr project is to generate, curate, interlink and publish economic data in LOD formats that are easily accessible and useful to the scientific community and to provide an user-friendly and objective layer of information that will enable citizens, journalists, business people and politicians to re-discover their own "stories" from data.

1. Introduction

The publicspending.gr project (PSGR) is a research initiative that has been initiated in 2011 by the Multimedia Technology Laboratory, School of Electrical and Computer Engineering at the National Technical University of Athens, in order to process and distribute linked open data concerning national expenditure.

After long discussions, presentations and hard work with diverse groups of people -ranging from the inceptors of the Clarity project[1] (see also [7, 8]) and post-graduate students that write their theses on PSGR to first time users of the application- we decided to enrich the application with advanced search features and domain level information in the fields of local government, education, physical persons, companies, health and energy.

We have completed the extensive process of data cleansing and we continue the linking with external datasets (i.e. product and activity classification schema alignment, geopolitical data, DBpedia and data related to the Greek economy). We have also initiated the creation of innovative visualizations based on mathematical network analysis.

Given the above, our initial two-fold goal remains unchanged: (a) generate, curate, interlink and publish economic data in LOD formats that are easily accessible by the scientific community and (b) provide an user-friendly and objective layer of information that will enable citizens, journalists, business people and politicians to re-discover their own "stories" from data. The remainder of the paper is as follows. In section 2 the data cleansing process is discussed, while section 3 briefly presents the linking with external datasets. Section 4 introduces the network statistics and visualizations of Greek public expenditures, while the last one discusses the way forward.

2. The data cleansing process

From the beginning of October 1st 2010, more than 30.000 public servants from all public organizations in Greece (including universities, hospitals etc.) upload in the Clarity web application every single decision, which is digitally signed and assigned a transaction unique number automatically by the system.

There is a series of errors or missing values in this process mainly due to the lack of appropriate restrictions during the data entry (e.g. VAT and CPV numbers). This is the reason behind the fact that many researchers leave unfinished their effort with the Clarity dataset. From the first moment, we believed that if we manage to handle the basic errors, then the unprecedented potential of this dataset would be revealed. We found out that the major problems in the dataset have been identified to be the **data format errors** (e.g. numbers as strings), **incorrect entries** (e.g. VAT or payee name) and **missing values** (mainly CPV codes).

Having pointed the major data format errors, a series of automatic error correction mechanisms have been developed by our part (using the Jena environment). Until the end of February 2013, some 200,000 decisions with wrong VAT entry and 4,000 with wrong CPV entries have been fixed (corresponding total amount of expenditure after these corrections reaches 10 billion euros).

Syntactic or semantic types of errors mainly drive incorrect entries. *Syntactic errors* on the VAT or payee name are corrected by calling the business registry service of the Tax Information System, which is operated by the General Secretariat of Information Systems (http://www.gsis.gr/). The service utilizes the Web Service Definition Language (WSDL), which is an XML format for describing service functionality. The querying is performed in the form of SOAP calls with the entity's VAT registration number as the reference key. The response contains metadata about the business entity, including contact details, activity descriptions, registration dates and current operational status.

Semantic errors are much harder to handle since they involve manual detection and correction. Since the launching of PSGR, it was noticed that for several physical entities, the amount of received payments was enormous and in the majority of the cases unreasonable. A closer investigation revealed a significant problem, which arose due to an

[1] "Cl@rity" Program: Every Government Decision on the Internet, http://diavgeia.gov.gr/en.

erroneous procedure that many users follow. Let us consider the case where a social security authority (payee) receives X euros in order to provide or cover the welfare allowance the for Y beneficiaries/citizens. In this example, each one of the Y beneficiaries has a distinct VAT registration number. Nevertheless, in most of the cases in the Clarity program, a sole physical entity that probably is the legal representative of the payee or even in some specific cases a random representative out of all beneficiaries (usually the one whose name is first in alphabetical order) is stated erroneously as the payee. This procedure occurs as a "solution", mainly because there is no multiple VAT registration number provision to such kind of payments. As a result, many physical persons along with their VAT registration numbers are "loaded" with huge payments, which are totally irrelevant with their income. Thus, in order to provide a solution, we followed a semi-automatic procedure that helped us clean the respective data. Through SPARQL queries we gathered in CSV format the top-2000 physical persons (records) who were labelled as payees from the beginning of the Clarity program up to the end of September 2012. Then for each separate record, we gathered their top-10 payments. The total amount of payments reached the level of 550M Euros.

Training Phase

Obviously, the manual check of such a vast amount of decisions is a prohibited and time-consuming procedure. So, we decided to split the work between five reviewers and manually check all top-50 records (10 records per reviewer). The number of decisions investigated by all reviewers was 342. Our target was to create a significant knowledge base over the most frequent CPV cases where the problem addressed appears more frequently. After the manual evaluation period for the top-50 records and their top-10 payments decisions, we ended up in the conclusion that two specific CPV clusters highly correspond to erroneously recorded cases between public/governmental organizations (as payers) and physical persons (as payees) as described in the beginning of this sub-section. The CPV clusters were the CPV codes that their first digits begin with 75 and 98. Such clusters appeared in 13 out of the 50 examined records. The total payment amount that corresponds to these two clusters was nearly 36.7M euros, which is 21.7% of the whole payments that appears in the training set.

Having examined one-by-one all payments of the cluster records, we constructed the following simple rule as an automatic data cleansing mechanism, where NewName appends Name in our initial CSV attribute schema and corresponds to the name of the payer entity in the xth payment decision, instead of the name of the physical name as it is recorded by the Clarity program.

The cleansing rule is: IF PaymentCpvCode={"98******" OR "75******"} THEN NewName(x)=PaymentPayerName(x)

Test-Evaluation Phase

After the manual training phase we proceeded with a statistically based test approach in order to evaluate how "quickly" and "easily" we can alleviate the problem, thus cleaning our data in a semi-automatic way.

Having in mind that {75******, 98******} CPV clusters are responsible for many errors; we matched those records from our CSV list for the payees between the top-51 and the top-500 position. We derived with 60 records that consisted of 439 payment decisions in total. The total corresponding payment amount was 26.8M Euros, which is nearly 14.4% of the whole amount of payments for the payees between the top-51 and the top-500 position (185.6M Euros). By performing a random statistical sampling with probability value equal to p=0.1 and after manual evaluation by the same reviewers employed in the training phase, we noticed that above than 93% of our testing cases (41 out of 44 randomly selected) can be recovered under the suggested cleansing rule.

The third type of errors in expenditure entries is the missing values in the CPV field. Our methodology has to do with the semi-automatic completion or correction suggestion of CPV field in spending decisions as these are announced through the Clarity project. Moreover, it was applied in the first top-20 organizations, which presented the larger total amount of payments, as this is derived from their top-200 payments in several payees.

For the selected organizations and for the top-200 spending decisions, we have manually suggested a valid CPV in case of no CPV appearance, or we have proposed a better CPV according to our opinion. These corrections were applied for decision between October of 2010 up to November 2012.

Reviewing single data entries is laborious and time consuming but the safest way to get a real grip on a dataset. Our effort to correct some thousands of payment entries with total value more than 40% of total payments has the following outcome

1. more meaningful statistics and visualizations,
2. productive proposals regarding the operation of Clarity project and
3. crowdsourcing the error correction mechanism.

We have presented our experience on the Clarity dataset in various meetings with the project management team in the Ministry of Administrative Reform. Most of our technical proposals have been already considered to be included in the next version of the program. The PSGR version that released on February 2013 introduces an error reporting form for every single payment entry and specific type of errors. Results are included in the PSGR dataset and are forwarded to the official authority that is responsible for the specific entry. In the near future, we plan to co-operate with the

National Center for Scientific Research (NCSR) - Demokritos (http//gov.insight.iit.demokritos.gr/) to include sentiment analysis in the incoming comments from users.

3. Linking with external datasets

Following the conversion of the PSGR dataset in RDF form, we have identified the following four areas that can act as linking points with external datasets (Figure 1):

1. product and activity classification schema alignment,
2. geopolitical data,
3. DBpedia and
4. data related to the Greek economy.

As has been mentioned, each payment is associated with a particular CPV (Common Procurement Vocabulary) code, which is a standardized classification adopted by the EU. Furthermore, each business agent is given a unique CPA (Classification of Products by Activity) code by the Greek taxation authority, which is available in our dataset. However, Procurement classifications do not follow a uniform standard. For instance, the European Union uses the Common Procurement Vocabulary (CPV) for business sectors, while the United States follows the North American Industry Classification System (NAICS) and the United Nations follows the Standard Products and Services Code (UNSPSC). For a comparative analysis of product classification schemata, see the work described in [4]. The MOLDEAS project [1] is an effort to align this plethora of classification systems and publish them as RDF with controlled semantics, following SKOS [6] modelling principles. URIs representing CPV and CPA codes in our dataset are linked to their corresponding CPV and CPA codes within the MOLDEAS dataset, thereby opening the way to spending comparisons on procurement categories that follow different classifications (e.g. data.gov).
The second area has to do with geopolitical information that is associated with the business register entries. Given that each business agent is associated with geospatial information (in the form of associations with postal code areas), we have identified and linked the relevant geopolitical resources (regions, regional units, municipalities, country) with their representations in DBpedia and Geonames. In particular, we take care of the mapping between postal codes and postal code areas (i.e. geopolitical areas uniquely defined by postal codes in a 1:1 fashion) and create postal code area resources, which we -in turn- link to geopolitical resources at different hierarchical levels. This makes possible the association of spending and business register data with resources that exist within DBpedia and Geonames and to form queries such as finding statistical information of spending by area, or spending by population become trivial. Furthermore, as we traverse the DBpedia graph we can associate the PSGR dataset with a large amount of external resource nodes with a linking path connecting them semantically, giving answers to potentially interesting queries.
Currently, we are building direct connections of the payment agents to el.dbpedia and DBpedia Live (e.g. Athens). The linking is done in the geo-information level. We do not link the authority, but the geopolitical region that the authority serves (distinction between the authority "Municipality of Athens" and the region that is governed by this authority). This procedure has been completed for all the national local government authorities (almost 400 organizations) and is on going for the payees coming from the private sector. They have been identified 1000 (17,7% in the total) private agents with Wikipedia lemmas. The fourth category of linking involves data that are related to the Greek economy.

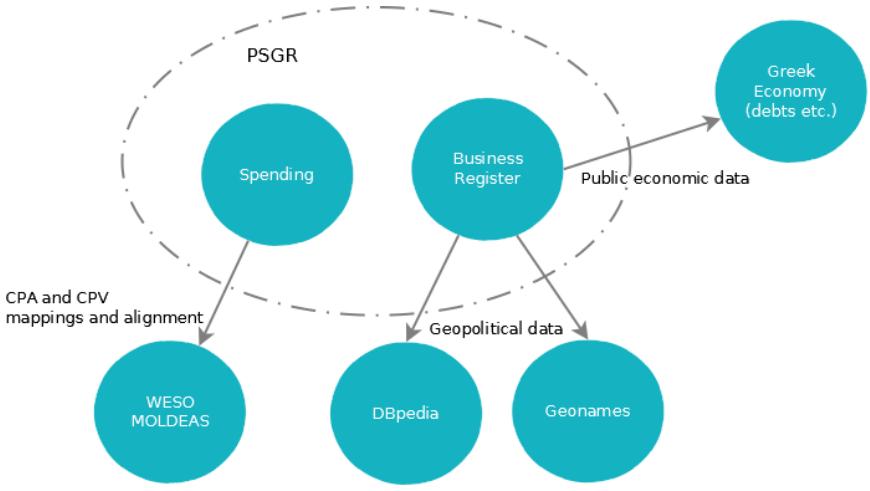

Figure 1: the links of PSGR with external datasets

The Greek Tax Authorities during 2012 have started to publish an updated list of the biggest debtors to the Greek government, including both physical and legal entities. A cross search between the sets of the payees and the debtors of Greek public concluded that 416 entities are both payees and debtors to the public (see the relevant SPARQL query results http://bit.ly/14hrhx8). In this context, the public e-procurement portal2 will offer data on the public tenders and contracts that can be interconnected to the PSGR dataset complementing the economic LOD [10], as an important part of the Web economy [9].

4. Data "storytelling": the case of network statistics on public expenditure

Currently, the most common tools for creating "stories" from data in science and journalism are infographics. In cases where big data are involved, infographics should be combined with sophisticated statistical methods. Anyway, the underlying trade-off between simplicity and informativeness remains in first place. We argue that visualizing the PSGR dataset as a payment network gets a good score on this trade-off. Despite the fact that mathematical network analysis has been successfully implemented in many types of economic networks [3], has not yet been applied in public spending data. The idea is simple as it relies on the fact that the data produced by PSGR represent a payment network. The network is formulated by the payments coming from public agents (payers) to payees (mainly private but could also be public). Nodes are either payers or payees, linked through a payment that is uniquely characterized by its amount, timestamp and category (CPV). For the descriptive statistics of the network refer to Table 1.

Descriptive statistics

Number of Nodes	115.870
Total Sum of Weights (euros)	6.920.573.743
Time period	18/10/2010-15/11/2012
Network Diameter	11
Average Weighted Degree (euros)	59.727
Average Path Length	4,5
Maximum Degree	7.779
Average Degree	3,2
Maximum In Degree	885
Average In Degree	1,6
Maximum Out Degree	7.775
Average Out Degree	1,6
Number of Nodes with 0 In Degree	1.277
Number of Nodes with 0 Out Degree (payees)	114.593
Number of Nodes with In & OutDegree	411
Number of Nodes with OutDegree (payers)	1.688

Table 1: descriptive statistics of the Greek public spending network

The network that has been produced by the Gephi software [2] has the following characteristics: (a) sparse (density<0.025), (b) strongly connected components 115.734 (nodes 99.75%, links 99.89%) and (c) weakly connected 86 (Nodes 0.25%, links 0.11%). The network is *connected*. Power law has been identified at the in-degree distribution by using linear regression for fitting (γ=1.77).

Node Importance

A centrality coefficient is a measure that captures the importance of a nodes or link's position in the network. There are local measures like degree centrality (in-out degree, weighted degree) and measures relative to the rest of the network such as betweeness and eigenvector centrality.

The degree centrality of a node is its degree. Nodes with more connections tend to have more power. In our case the degree of a node (agent) in the graph represents the number of the agents that it transacts with. The higher the degree is for an agent the bigger is the importance of the agent for the network as it either acts as a receiver of services or goods (out degree), or acts as a supplier of services or goods (in degree). Given the nature of the network (public spending), nodes with high out degree are expected to be public agents and nodes with high in degree private sector agents.

Top brokers

Some 411 public agents act as brokers (diffuse money into the network) by both receiving and paying money to other agents. These agents are interconnecting the graph and are characterized by high score in betweenness centrality.

Betweenness centrality describes the extent to which a particular node lies between other network nodes. As a measure, it takes into account factors such as the connectivity of the node's neighbors, giving a higher value for nodes, which bridge clusters. In economic relationships, betweeness is an index of local importance of an agent in payments.

2 Its pilot phase started in February 2013 in http//www.eprocurement.gov.gr.

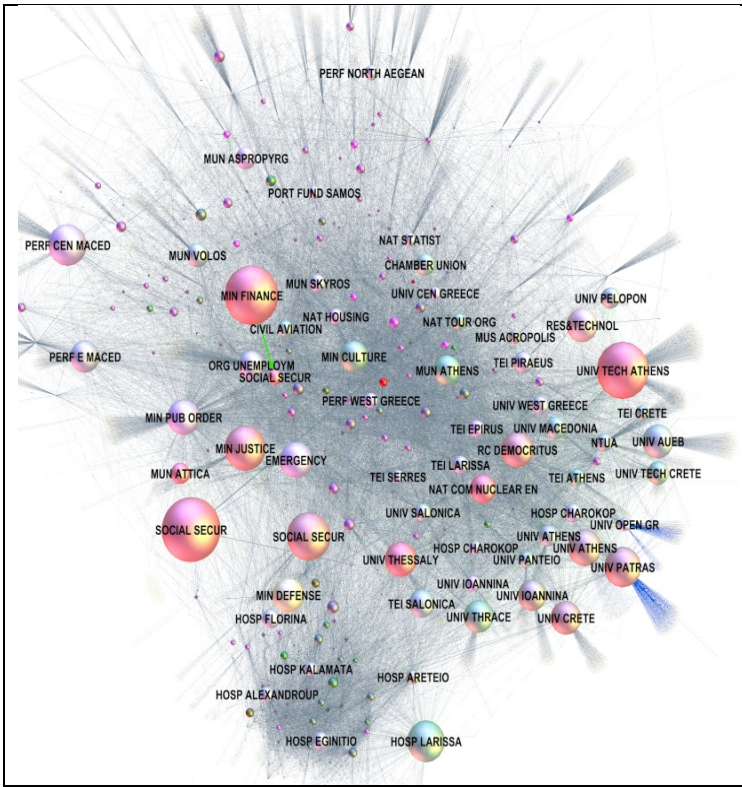

Figure 2: agents that act as brokers in the Greek public spending network
Node size: Betweeness Centrality,
Node color: Hub Ranking

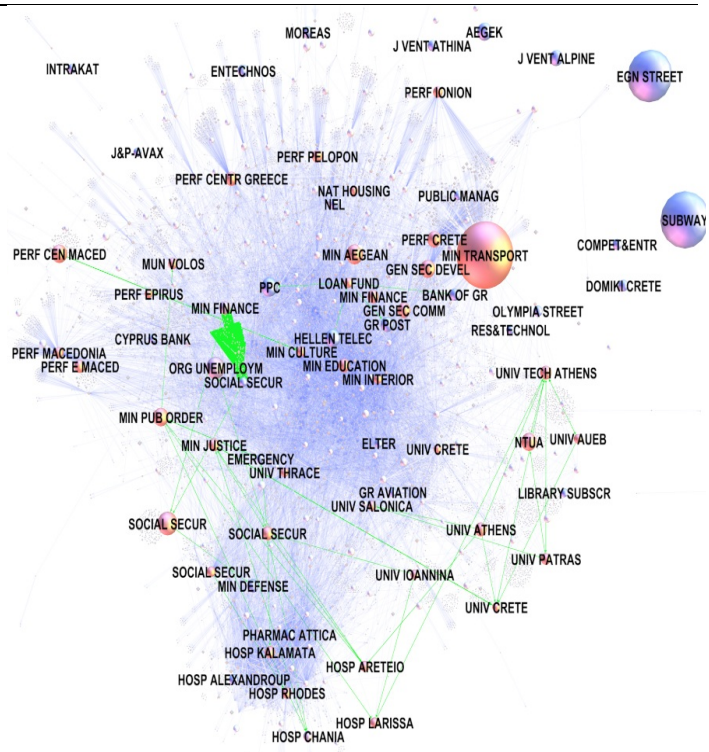

Figure 3: top payment agents in the Greek public spending network
Node size: Weighted Degree Centrality,
Node color: Degree Rank

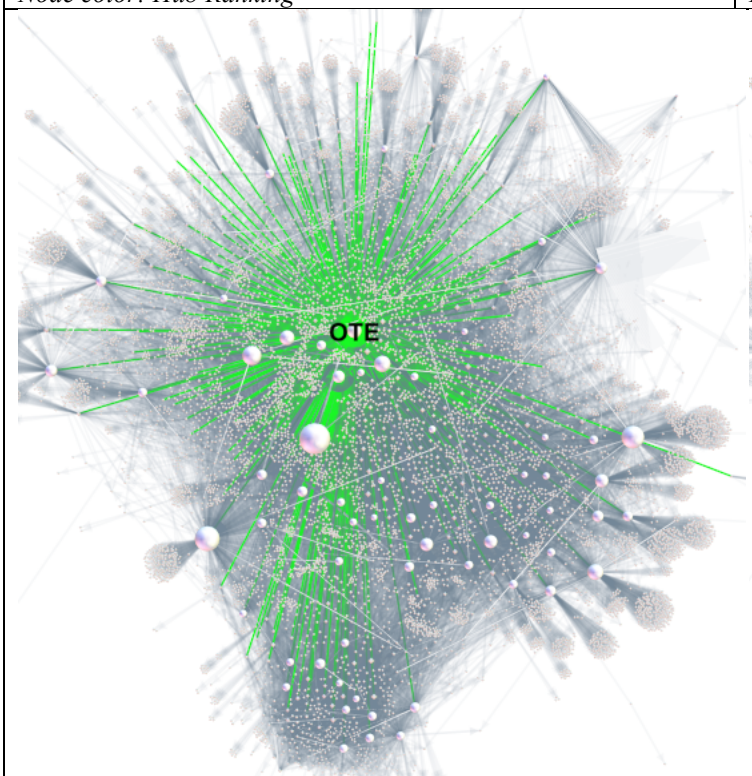

Figure 4: the Hellenic Telecommunication Organization (OTE) is the agent with the highest in-degree in the Greek public spending network (*green arrows=incoming payment).*

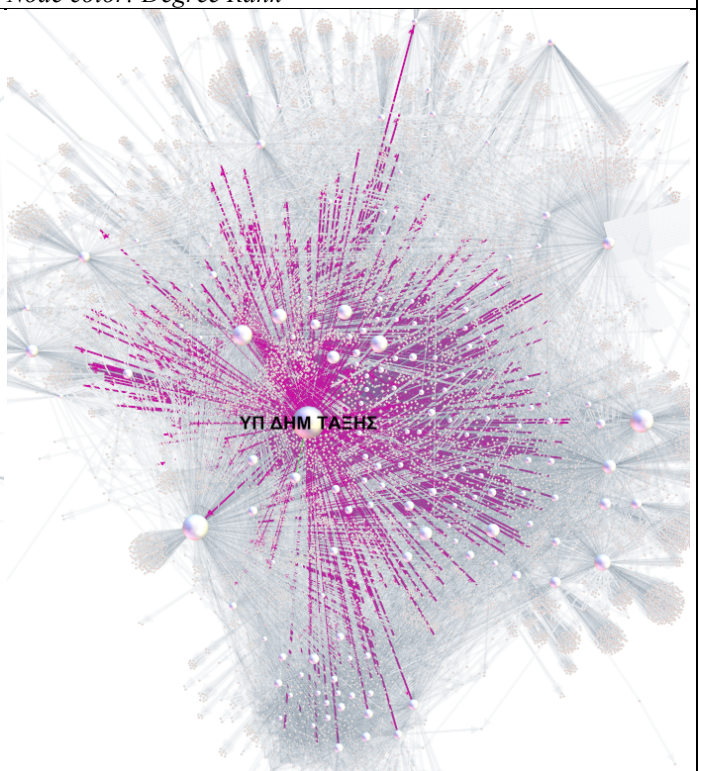

Figure 5: the Ministry of Public Order is the agent with the highest out-degree in the Greek public spending network *(red arrows=outgoing payment).*

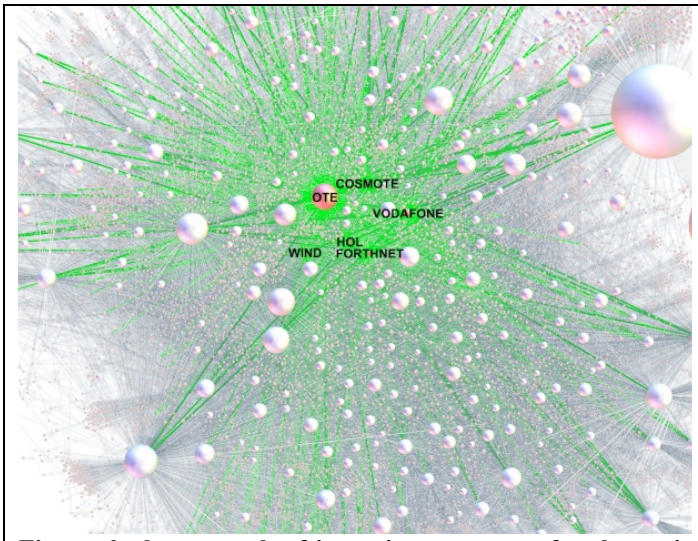

Figure 6: the network of incoming payments for the major telecommunication players in the Greek public spending network.

Our objective is to capture interesting relations among underlying agents through graph visualizations. Results can be summarized as follows:

1. There two major projects in Greece the Subway in Athens and Thessaloniki and Egnatia Runway in North Greece (Figure 3).
2. The major constructions projects in the country are funded through the General Secretariat of Public Works.
3. The Ministry of Public Order was involved in significant construction works in regard to other ministries or regional authorities (Figure 5).
4. Universities and research institutes are important hubs in the network maintaining a wide network of payees that offer a variety of services for them and are significant contributors to the dispersion of funds. This also applies to Regional Authorities in local scale (Figure 2).
5. Pension Funds, Labour Office and other social security institutes realize a significant amount of payments for services and are important agents in the network, independently of their actual spending for pensions or social security allowances (Figure 2).
6. Hospitals are also important hubs (spenders) in the network. They spend mainly on special services, drugs and materials specialized for medical use and therefore a cloud of companies (payees) is gathered around them offering the needed (Figure 2).
7. Public utilities, such as the Hellenic Telecommunication Organization (OTE), are the payees with the highest indegree in the Greek public spending network (Figure 4).
8. OTE is the winner in the competition for customers in the public sector (Figure 6).

5. Discussion

The provision of linked open data in public spending domain will enable a thorough, easy and updated online access to who, when, why and what a public institution spends. Furthermore, the generic information about public expenditure can initiate a vast range of analysis from mathematical networks to data journalism. These include data visualizations, economic and statistical analysis and policy assessment. Beyond transparency, economic LOD provides an extra tool to governments for implementing advanced budget monitoring processes and multiple and more complex criteria for eligibility in e-procurement with minimum costs. In the business side, economic LOD can be dynamically integrated in business information systems for stimulating dynamic resource allocation with lower cost and wider application range [5]. This type of integration will benefit more the SMEs, since they lack advanced business intelligence systems and sufficient R&D budgets.

6. References

[1] ALVAREZ, J. and LABRA, J. 2012. Towards a pan-european e-procurement platform to aggregate, publish and search public procurement notices powered by Linked Open Data: the MOLDEAS approach. *International Journal of Software Engineering and Knowledge Engineering*. 22, 3 (2012), 365–383.

[2] Bastian, M. et al. 2009. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media* (2009).

[3] Jackson, M. 2008. *Social and economic networks*. Princeton Univ Pr.

[4] Maniatopoulos, J. and Leukel, J. 2005. A comparative analysis of product classification in public vs. private e-procurement. *Electronic Journal of e-Government*. 3, 4 (2005), 201–212.

[5] Meimaris, M. and Vafopoulos, M. 2012. Knowledge-Based Semantification of Business Communications in ERP Environments. *Engineering the Semantic Enterprise Workshop in the 13th International Conference on Web Information System Engineering (WISE 2012)* (2012).

[6] Skos simple knowledge organization system primer: 2008. *http://www.w3.org/TR/2008/WD-skos-primer-20080829/*. Accessed: 2013-03-02.

[7] Vafopoulos, M. et al. 2012. Intelligent and semantic real-time process of the Greek LOD for enhancing citizen awareness in public expenditures. *The 13th International Conference on Web Information System Engineering (WISE 2012)* (2012).

[8] Vafopoulos, M. et al. 2012. Publicspending. gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives. *USING OPEN DATA: policy modeling, citizen empowerment, data journalism* (2012).
[9] Vafopoulos, M. 2011. The Web economy: goods, users, models and policies. *Foundations and Trends® in Web Science*. 3, 1-2 (2011), 1–136.
[10] Vafopoulos, M. and Meimaris, M. 2012. Weaving the economic Linked Open Data. *7th International workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (2012).