QLectives – Socially Intelligent Systems for Quality

Project no. 231200

Instrument: Large-scale integrating project (IP)

Programme: FP7-ICT

Deliverable D3.2.1

[D3.2.1 Datasets guide and manual – external datasets]

Submission date: 2012-02-29

Start date of project: 2009-03-01 **Duration: 48 months**

Organisation name of lead contractor for this deliverable: University of Warsaw

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	**x**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document information

1.1 Author(s)

Author	Organisation	E-mail
Michal Ziembowicz	University of Warsaw	ziembowicz@gmail.com

1.2 Other contributors

Name	Organisation	E-mail
Klara Łucznik	University of Warsaw	klara.lucznik@gmail.com
Tamas Vinko	TU Delft	tamas tamas.vinko@gmail.com
Matus Medo	University of Fribourg	matus.medo@unifr.ch

1.3 Document history

Version#	Date	Change
V0.1	01 December, 2011	First draft
V0.5	10 January, 2012	First final version
V1.0	29 February, 2012	Approved version to be submitted to EU

1.4 Document data

Keywords	external datasets, batch0, data collection
Editor address data	ziembowicz@gmail.com
Delivery date	29 February, 2012

1.5 Distribution list

Date	Issue	E-mail
29.02. 2012	**Consortium members**	QLECTIVES@LIST.SURREY.AC.UK
20.03.2012	**Project officer**	Jose.FERNANDEZ-VILLACANAS@ec.europa.eu
20.03.2012	**EC archive**	INFSO-ICT-231200@ec.europa.eu

QLectives Consortium

This document is part of a research project funded by the ICT Programme of the Commission of the European Communities as grant number ICT-2009-231200

.

University of Surrey
(Coordinator)
Department of Sociology/Centre
for Research in Social Simulation
Guildford GU2 7XH
Surrey
United Kingdom
Contact person: Prof. Nigel Gilbert
E-mail: n.gilbert@surrey.ac.uk

Technical University of Delft
Department of Software
Technology
Delft, 2628 CN
Netherlands
Contact Person: Dr Johan
Pouwelse
E-mail: j.a.pouwelse@tudelft.nl

ETH Zurich
Chair of Sociology, in particular
Modelling and Simulation,
Zurich, CH-8092
Switzerland
Contact person: Prof. Dirk Helbing
E-mail: dhelbing@ethz.ch

University of Szeged
MTA-SZTE Research Group on
Artificial Intelligence
Szeged 6720, Hungary
Contact person: Dr Mark Jelasity
E-mail: jelasity@inf.u-szeged.hu

University of Fribourg
Department of Physics
Fribourg 1700
Switzerland
Contact person: Prof. Yi-Cheng
Zhang
E-mail: yi-cheng.zhang@unifr.ch

University of Warsaw
Faculty of Psychology
Warsaw 00927, Poland
Contact Person: Prof. Andrzej
Nowak
E-mail: nowak@fau.edu

Centre National de la Recherche Scientifique, CNRS
Paris 75006,
France
Contact person: Dr. Camille
ROTH
E-mail:
camille.roth@polytechnique.edu

Institut für Rundfunktechnik GmbH
Munich 80939
Germany
Contact person: Dr. Christoph
Dosch
E-mail: dosch@irt.de

QLectives introduction

QLectives is a project bringing together top social modelers, peer-to-peer engineers and physicists to design and deploy next generation self-organising socially intelligent information systems. The project aims to combine three recent trends within information systems:

- **Social networks** - in which people link to others over the Internet to gain value and facilitate collaboration
- **Peer production** - in which people collectively produce informational products and experiences without traditional hierarchies or market incentives
- **Peer-to-Peer systems** - in which software clients running on user machines distribute media and other information without a central server or administrative control

QLectives aims to bring these together to form Quality Collectives, i.e. functional decentralised communities that self-organise and self-maintain for the benefit of the people who comprise them. We aim to generate theory at the social level, design algorithms and deploy prototypes targeted towards two application domains:

- **QMedia** - an interactive peer-to-peer media distribution system (including live streaming), providing fully distributed social filtering and recommendation for quality
- **QScience** - a distributed platform for scientists allowing them to locate or form new communities and quality reviewing mechanisms, which are transparent and promote

The approach of the QLectives project is unique in that it brings together a highly inter-disciplinary team applied to specific real world problems. The project applies a scientific approach to research by formulating theories, applying them to real systems and then performing detailed measurements of system and user behaviour to validate or modify our theories if necessary. The two applications will be based on two existing user communities comprising several thousand people - so-called "Living labs", media sharing community tribler.org; and the scientific collaboration forum EconoPhysics.

Table of contents

Introduction

The QLectives project's work is based on the process of data collection and analysis. In each year a new portion (called a "batch") of data is fed from Stream 3 and 4 into Streams 1 and 2. Whereas data used in the years 2 and 3 comes from the Living Labs (QMedia and QScience), in the first year the data from QLectives was not yet available and therefore external sets were used. The aim of deliverable D 3.2.1 is to prepare a manual for those external datasets collected in Batch 0.

The databases were selected during an early meeting (May 2009, Delft) by the partners involved in Streams 3 and 4. The criteria for including a dataset into Batch 0 were based on its potential usability for research. It focused on the social aspect of the data so databases describing human agents' and social activities data were chosen.

Three main categories can be found in the set: social networks, peer-to-peer and online communities. The first category concentrates on the structure of networks of human interactions: the structure of a network based on trust relations in case of PGP, and the structure produced in the process of communication in case of Emailing network. Another area covers the second category of datasets related to peer-to-peer (p2p) files exchange. Most p2p research concentrates on the level of "swarms" disregarding the impact from individual users. Here, in contrast, the two datasets (FileList and BitSoup) were introduced that were gathered on the peer level. Finally the last category comprises the different approaches to online communities. A hypergraph theory is used to model networks of experts. The evolution of communities is studied in the Wiki and FlickR environments.

Each dataset is presented with a short description and an overview of measures it provides. Further there are links to sample data files and references to articles using these data. Also in each case a way of obtaining the whole dataset is presented.

1. PGP contacts graph

Description:

The dataset was used in the research paper by Boguñá, Pastor-Satorras, Díaz-Guilera & Arenas (2004).

The data consists of a list of edges of a snapshot of Pretty-Good-Privacy algorithm's web of trust as it was on July 2001. Only bidirectional signatures (i.e., peers who have mutually signed their keys) have been considered. This filtering process guarantees mutual knowledge between connected peers and makes the PGP network a reliable proxy of the underlying social network. The complete undirected network is composed by 57243 vertices with an average degree <k>=2.16. The giant component (GC) of this network, i.e., the largest connected subnetwork, (which is the data actually provided) comprises 10680 vertices and its average degree is <k>GC=4.55.

The interest of the PGP network is twofold: First, it is a web based on trust, and the comprehension of trust networks is, nowadays, crucial to understand the complexity of the information society. Second, unlike collaboration networks, this web is one of the largest reported nonbipartite graphs one can build from large databases in social sciences. The consideration of this web of trust as a benchmark for the evaluation of the proposed social model is, thus, fully justified.

Available measurements:

Social network structure:

- network size: 57243 vertices,
- average degree <k> = 2.16,
- giant component size: 10 680 vertices,
- giant component average degree: $\langle k \rangle_{GC}$=4.55 .

Availability:

A sample of data can be found here: http://www.QLectives.eu/wiki/images/1/13/PGP.sql , the complete dataset can be obtained from ETH Zurich.

References:

- Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., Arenas (2004) A. Models of social networks based on social distance attachment. *Phys Rev* E 70: 056122

2. URV email traffic graph

Description:

The dataset was used in the research paper by Guimera, Danon, Diaz-Guilera, Giralt & Arenas (2003). The data consists of a list of edges of the network formed by e-mail interchanges between members of the University Rovira i Virgili (Tarragona, Spain). In this network, each email address is a node and a link between two of them implies an email communication. Only email communications that are bidirectional (A sent an email to B and vice-versa) and sent to less than 50 recipients have been considered. With these restrictions, the network is an undirected graph. The dataset corresponds to the giant component of such a network, which presents a clustering coefficient C=0.254 and an average shortest path length d=3.606.

Available measurements:

Network structure and communication tracks:

- number of nodes (in the giant component): 1133
- clustering coefficient C = 0.254,
- average shortest path length d = 53.606

Availability:

A sample of data can be found here: http://www.QLectives.eu/wiki/images/2/27/Email.sql , the complete dataset can be obtained from ETH Zurich.

References:

- Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A., (2003) Self-similar community structure in a network of human interactions. *Physical Review*, Vol. E 68, Nr. 065103.

3. FileList.org

Description:
The FileList.org is a private (closed) BitTorrent community. Its tracker was scraped from 9 December 2005 until 4 April 2006, when FileList changed the setup of their website and removed the online swarm member lists. As the scrape of the whole community was done in every 6 minutes, this is a unique dataset capturing the dynamic trends of interactions in BitTorrent communities.

Available measurements:
We have collected a total of 80 GB of compressed data, which contains statistics about the members (more than 90,000 peers) of over 3,000 download swarms. From the pure HTML files a database containing the BitTorrent sessions was created. In this database the available measurements are:

- session-length (sec);
- upload and download, that give the amount of uploaded and downloaded data during the session (KB);
- seeding-length, that is, the amount of time spent online after the file has been completely downloaded (sec);
- seeded, the amount of data uploaded during seeding (KB);
- up-speed and down-speed, that are calculated by taking the maximum of the download or upload speed, respectively, as reported by the tracker over the observation points of the session (KB/sec);
- and an indicator of whether the peer was behind a firewall or not.

Availability:
A sample of the dataset can be found at http://www.QLectives.eu/wiki/images/e/e6/Tom-data.sql , and complete datasets can be obtained from Johan Pouwelse (TU Delft).

References:
This database has been used in many studies:

- Csernai, K., Jelasity, M., Pouwelse, J. and Vinkó, T. (2011) *Modeling unconnectable peers in private BitTorrent communities*. In Proceedings of MSOP2P 2011. IEEE Computer Society.
- Capota, M. Andrade, N., Vinkó, T., Santos, F. R., Pouwelse, J. and Epema, D. (2011) *Inter-swarm resource allocation in BitTorrent communities*. In Proceedings of the 11th IEEE International Conference on Peer-to-Peer Computing (P2P11). IEEE. pp. 303-309.
- Rahman, R., Hales, D., Vinko, T., Pouwelse, J. and Sips, H. (2010) *No more crash or crunch: sustainable credit dynamics in a P2P community*. International Conference on High Performance Computing & Simulation (HPCS 2010), Caen, France, 2010.
- Ormándi, R., Hegedűs, I., Csernai, K. and Jelasity, M. (2010) *Towards inferring ratings from user behavior in BitTorrent communities*. In Proceedings of the 6th International Workshop on Collaborative Peer-to-Peer Systems (COPS) at WETICE'10, pp. 217–222. IEEE Computer Society, 2010. Best Paper Award. (doi:10.1109/WETICE.2010.41)
- and also in the ongoing work described in the deliverable D1.4.1.

4. FileList and BitSoup (Seeders' resource allocation in two BitTorrent communities)

Description:
Measurements of current and possible resource allocations in two BitTorrent communities (FileList and BitSoup).

Available measurements:
The datasets contain resource allocation (seeding and leeching) of ~90,000 users in both communities, resulting in ~76,000 and 32,000 active sessions in BitSoup and FileList, respectively.

Allocations are measured at instants. There are 100 such measurements for Bitsoup and 88 for Filelist. For each timestamp, there are three xml files: one which expresses the numbers of seeders and leechers in each torrent, and has no seeder unallocated, and two files where torrents are listed with no seeder in them and seeders are listed with their capacity not allocated and with the list of torrents they can seed at the measurement instant. The two files represent different estimation strategies for files currently owned by each seeder.

Availability:
A sample of the dataset can be found here http://QLectives.eu/wiki/datastore/seeders_allocation.tar.gz , and complete datasets can be obtained from Johan Pouwelse (TU Delft).

References:
This datafiles were used in the paper:

- Capota, M. Andrade, N., Vinkó, T., Santos, F. R., Pouwelse, J. and Epema, D. (2011) *Inter-swarm resource allocation in BitTorrent communities*. In Proceedings of the 11th IEEE International Conference on Peer-to-Peer Computing (P2P11). IEEE. pp. 303-309

5. Epistemic hypergraphs

Description:

These datasets consist of data about scientific collaboration. They reveal a large part of the underlying collaboration activity: temporal information on teams, gathering agents and the topics they work on, assuming that topics are described by the terms (lemmas/n- grams) used in paper abstracts. For each database, we focus on a set of no more than a hundred of relevant terms, selected with the help of an expert of the corresponding field and are such that they appropriately cover the most significant topics of each field.

Fields are defined either from a semantic perspective (using e.g. field names) or from a social perspective (using e.g. scientific assemblies), and involving both large and small communities:

1. Embryologists working within a given and well-determined subfield—the zebrafish, on a period of 20 years (1985–2004). Data was extracted from the publicly available database Medline, which eventually yields a dataset of 6,145 articles (13,084 authors, 71 word classes).
2. Scientists working on rabies from the same kind of MedLine extraction as for zebrafish embryologists—the observed period spans from 1985 to 2007. This ends up with 4,648 events (9,684 authors, 70 word classes).
3. Scientific committee members for JEMRA meetings (FAO/WHO experts): this dataset includes the publications of an initial set of 168 scientists involved in these meetings, gathered from 1985 to 2007. This leads to 5,893 papers (15,375 authors, 69 word classes).
4. Scientific committee members for JECFA meetings (FAO/WHO experts): similarly, publications of an initial set of 178 scientists are gathered from 1985 to 2007. This ends up with 8 685 papers (21 195 authors, 85 word classes).

This dataset was used to introduce hypernetworks (hypergraphs) rather than networks (graphs) as a relevant level to model collectives, exploring the theoretical implications of hypergraphs in network analysis (especially in terms of observed patterns in random hypergraph models) and studying the intertwinement between social, semantic and quality-related aspects of group formation. This aspect sheds light on an important issue of QScience: the relationship between various notions of quality and underlying social structural patterns. An article (on the likelihood of some patterns of team formation) has already been published (Taramasco, C., Cointet, J.-P. & Roth, C. (2010) Academic team formation as evolving hypergraphs. *Scientometrics*, 85(3):721–740), another one (on the relationship between team patterns and team impact) is in preparation as of Feb 2012.

Available measurements:

This dataset consists of evolving hypergraphs: a hypergraph is a generalization of a graph where links can gather any number of nodes, i.e., they are isomorphic to events (gathering groups of items). For each of the four datasets, there are around half a dozen events, with a little below a hundred concepts and between ten and twenty thousand individuals. On top of the above-described numbers of social and semantic nodes and events, we measured:

- the hypergraphic repetition ratio for concepts and individuals,
- the expertise ratio (presence of individuals who were already present in past timesteps);
- the propensity of group formation with respect to hypergraphic repetition ratio (agents and concepts), and with respect to expertise ratio;
- and several correlations between these variables.

Availability:
Samples of the datasets can be found here http://camille.roth.free.fr/software.php , and complete datasets can be obtained from Camille Roth (CNRS).

References:

- Morphogenesis of epistemic networks: a case study, Camille Roth. ESSA 4th Conf. European Social Simulation Association, Toulouse, France, Sep 2007

6. Wiki

Description:

The dataset consists of dynamic data on the development of a sample of more than ten 11,500 MediaWiki-based wikis selected by the s23.org website (see http://s23.org/wikistats/largest_html.php) over the period of August 2007–April 2008.

The dataset has been used for a case study on some factors likely to account for the diverse success of wiki platforms, in terms of technical, social and structural features. In this context, we understand "viability" as dynamic sustainability of both population and quality content: in other words, a viable wiki should be able to grow in terms of articles and users in such a way that the whole content can be maintained by a sufficient number of users. As content-based online communities, wikis mainly evolve in two dimensions: (a) contributors, who may or may not constitute an active community; and (b) pages, which may or may not amount to authoritative or useful content. The results suggest that different structural and governance-related factors have significant correlation with – and plausibly, in some cases, effect on – the content and population dynamics of a wiki.

Available measurements:

11,500 MediaWiki-based wikis monitored on 4 dimensions for 9 months on a daily basis, resulting in a database of 322 MB (compressed). The data consists of features that are publicly available from MediaWiki platforms: number of users, pages, administrators and edits. The data has been collected on a daily basis, providing precise longitudinal data. (It is straightforward to deduce growth rates from the longitudinal analysis of the database.)

The dataset has been further completed by a single crawl aimed at determining whether wikis were editable without having to sign up (permission to anonymously edit), which occurred at the end of the monitoring period (April 2008).

Availability:

Samples of datasets can be found here: http://QLectives.eu/wiki/datastore/wikistats.tgz, and complete datasets can be obtained from Camille Roth (CNRS).

References:

- Roth, C.; Taraborelli, D.; & Gilbert, N. (2008) Measuring wiki viability: An empirical assessment of the social dynamics of a large sample of wikis. In Proceedings of the 2008 International Wiki Symposium.
- Taraborelli, D., Roth, C., and Gilbert, N. (2008) Measuring wiki viability (II). Towards a standard framework for tracking content-based online communities. Tech. rep., 2008
- Taraborelli, D., Roth, C. (2011) Viable Web Communities: Two Case Studies, in *Viability and Resilience of Complex Systems*, eds. G. Deffuant and N. Gilbert; Springer. p. 75-105.

7. Flickr Groups

Description:

Daily snapshots of groups since their registration to the service were collected using a web service written by Dario Taraborelli from 2008 to 2009. For a subset of 9360 groups, more fine-grained supplementary data collected between June and July 2009 are available including: (i) two snapshots of the unique IDs of the entire population of each group, and (ii) two snapshots of the unique group ID of the affiliations of each group member. In 2010, a further dataset was collected for 500 groups of similar size (100-140 members) relating to their photo contribution and commenting activity.

From analyses of the data, the main factors affecting the growth of a group were found to be its size (the greater the number of members, the faster the growth rate; this is consistent with herding behaviour) and cohesiveness (the greater the density of links between members, the slower its growth). Surprisingly, governance and moderation had very little impact. (Further details can be found in Taraborelli & Roth (2011))

From analyses of the small set of similar sized groups, we found support for the hypothesis that different members play different roles in the group, with some being highly committed contributors of photos, while others act as disseminators through having more social links or belonging to more other groups. (Further details can be found in Chen, C-C and Roth, C. (2010)).

Available Measurements:

14500+ groups (90MB). Finer grain snapshots for 9360 groups. Member-centric data for 500 groups.
For all groups:

- Number of users;
- Number of admins;
- Number of photos;
- Throttling type;
- Privacy type;
- Moderation type.

For the finer grain snapshots of 9360 groups:

- Average (directed) user degree;
- Average of clustering coefficient;
- Proportion of reciprocated links;
- Average membership spread (number of other group affiliations per member);
- Group membership turnover between the two snapshots.

For the 500 groups of similar size:

- (Directed) degree for each member;
- Number of photos contributed by each member;
- Number of groups member belongs to;
- Number of comments contributed by each member.

Availability:

Samples of datasets can be found here:

- http://www.QLectives.eu/wiki/images/4/40/Flickr_master_sample.txt ,
- http://www.QLectives.eu/wiki/images/4/40/Flickr_group_daily_sample.txt ,
- http://www.QLectives.eu/wiki/images/d/df/Flickr_master_sn_sample.txt,

and complete datasets can be obtained from Camille Roth (CNRS).

References:

- Chen, C-C and Roth, C. (2010) Motivations and Constraints to Member Engagament in Flickr Groups. QTESO workshop at the 4th IEEE SASO conference
- Taraborelli, D. and Roth, C. (2011) Viable Web Communities: Two case studies. In *Viability and Resilience of Complex Systems* ed. G. Deffuant and N. Gilbert, Springer