enhanced COntent distribUtion with Social INformation

www.ict-ecousin.eu

Deliverable D3.2

Initial Release of Measurement and Prediction Software (Prototype Deliverable)

Public deliverable, Version 1.0, 30 April 2014

Authors	
Orange	Ali Gouta, Yannick Le Louedec
IMDEA	Miriam Marciel, Joerg Widmer
TSP	Reza Farahbakhsh, Angel Cuevas, Xiao Han, Noel Crespi
TUD	Fabian Kaup, Julius Rückert, Tamara Knierim, Christian Koch, David Hausheer (Editor)
UCAM	Eiko Yoneki
UC3M	Roberto Gonzalez, Ruben Cuevas, Juan Miguel Carrascosa
Reviewers	Roberto Gonzalez, Fabio Mondin

Abstract

Deliverable D3.2 is a prototype deliverable which includes the first release of measurement and prediction software in eCOUSIN. D3.2 also provides initial datasets gathered with these tools. This report is an accompanying document to the actual prototype deliverable that is provided as an archive file containing the actual software tools and datasets. Specifically, D3.2 provides initial implementations of efficient algorithms for crawling, monitoring, and data gathering in social-based content centric infrastructures. In more detail, this includes software tools for large scale crawling and measurement of OSNs, for measurement of content distribution and content portals, as well as for monitoring of OSN and content distribution traffic in operational networks.

Project funded by the European Union under the
Information and Communication Technologies FP7 Cooperation Programme
Grant Agreement number 318398

EXECUTIVE SUMMARY

Deliverable D3.2 is a prototype deliverable, thus the actual deliverable is an archive file containing software tools and datasets. The aim of this document is to provide an overview on these prototype tools and datasets.

Specifically, Deliverable D3.2 provides the initial release of software tools for data collection, measurement, and prediction in social-based content centric infrastructures. Additionally, D3.2 includes also an initial set of traces collected with these tools.

Accordingly, the main results of Deliverable 3.2 include:

- A first release of measurement and prediction software (Section 1). This includes software tools for large scale crawling and measurement of OSNs such as Facebook, Twitter, and Google+, as well as software tools for measurement of content distribution and content portals such as YouTube, BitTorrent, and BTLive. Furthermore, tools for monitoring OSN and content distribution traffic in operational networks are described here as well. This section is completed with a simulator for content prediction.

- An initial set of data and traces (Section 2). This includes datasets that have been gathered with the above tools, specifically datasets of OSNs such as Facebook and Google+, as well as datasets of content distribution and content portals such as YouTube, BitTorrent, and BTLive. The section is completed with a description of traces from operational networks and with eye tracking and EEG data for modelling social cascades.

TABLE OF CONTENTS

1. MEASUREMENT AND PREDICTION SOFTWARE

This section provides an outline of the first release of measurement and prediction software. This includes software tools for large scale crawling and measurement of OSNs such as Facebook, Twitter, and Google+, as well as software tools for measurement of content distribution and content portals such as YouTube, BitTorrent, and BTLive. Furthermore, tools for monitoring operational networks are described here as well. The section is completed with a simulator for content prediction.

A detailed overview on the different software tools with their publication type and dissemination level is given below. (Public: Available publicly; Restricted/Confidential: Available upon request only).

	Software Tools	Publication Type	Dissemination Level
Facebook	Facebook Brands Crawling (Public)	Open Source	Public
	Facebook Brands Crawling (Restricted)	Open Source	Restricted
	Passive Facebook Measurement Software (Tracing App/SonNet)	Binary	Public
	Facebook Social Aggregator	Open Source	Public
Twitter	Twitter Locality	Open Source	Public
	Twitter Crawler (Tweet + Trends)	Open Source	Public
Google+ Crawler		Open Source	Public
YouTube Crawler		Open Source	Public
BT	BitTorrent Macroscopic Crawler	Open Source	Public
	BitTorrent Microscopic Crawler	Open Source	Confidential
BTLive	BTLive Wireshark Plugin	Binary	Public
	BTLive Web Crawler	Open Source	Public
Mobile Bandwidth Measurement App (NetworkCoverage)		Binary	Public
Simulator for Content Prediction		Open Source	Restricted

1.1 Facebook

1.1.1 Facebook Brands Crawling (Public)

Directory	Software/public/TSP_Facebook-Brands-Crawling-Public.rar		
Collected Dataset(s)	-		
Software Description	This tool is able to collect the general information of Facebook fan pages.		
Measured Parameters	This tool is able to collect the following data: brands' name, brands' category (which is selected by fan owners in the creation time and is available in their Fan page), number of likes, number of people talking about the pages.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Python	**Supported Operating Systems**	Window and Linux are tested.
Software License	GPL	**Dissemination Level**	Public
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities
Official URL	-		
Papers Published	-		
Software Installation	Brands_list.txt is input of the tool and includes Id and name of the Fan pages that are targeted to be collected.		

1.1.2 Facebook Brands Crawling (Restricted)

Directory	Contact Reza Farahbakhsh (reza.farahbakhsh@it-sudparis.eu) or Angel Cuevas Rumin (angel.cuevas_rumin@it-sudparis.eu)		
Collected Dataset(s)	Facebook Brands Data		
Software Description	This tool will be able to collect information from posts made in FB Pages both for Fan pages and regular users (only public posts). The tool is under development.		
Measured Parameters	For all published posts in a Facebook fan page and all public posts inside a regular user wall page, this tool is able to collect: Timestamp of creation and modification, type of posts (video, status, photo, etc.), number of likes and number of comments and number of shares.		
Measurement	Active	**Measurement**	Non-cooperative

Class		**Environment**	
Programming Language	Python	**Supported Operating Systems**	Window and Linux are tested.
Software License	GPL	**Dissemination Level**	Restricted
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities (upon acceptance of TSP)
Official URL	-		
Papers Published	-		
Software Installation	Brands_list.txt is input of the tool and includes Id and name of the Fan pages that we are targeting to be collected.		

1.1.3 Passive Facebook Measurement Software (Tracing App/SonNet)

Directory	Software/public/TUD_TracingApp.zip		
Collected Dataset(s)	Not yet ready for publication		
Software Description	SonNet is a Facebook client for the investigation of the use of video and other media content via online social networks. The app is used in the context of a user study at TUD. The aim of the investigations is to make predictions about access to content on the user's news feed, so as to design a pre-fetching mechanism for offloading the mobile data network. In context of the study, SonNet collects anonymized structural and content-related data (with prior permission of the user). By anonymizing we ensure that the app does not process personal data of the user. At the end of the study, the user is requested to transfer the data collected to evaluate it on our statistics server. Here the data is aggregated further and distorted, so that the then used data for processing does not allow any conclusions about the participating users.		
Measured Parameters	The app collects information about the content that a user sees on his personal news feed in Facebook. General statistics are collected, such as the distribution of posts among content types (i.e. video, picture, text, etc.), as well as social properties of the posts. Here, for example, information on the number of likes, comments, and shares of all posts are traced.		
Measurement Class	Passive	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	Android 4.0 and newer
Software License	The binary might be freely used and shared as it is.	**Dissemination Level**	Public

Publication Type	Binary	**Requirement to obtain the software**	None
Official URL	https://play.google.com/store/apps/details?id=de.tudarmstadt.kom.sonnet		
Papers Published	-		
Software Installation	For ease of simple installation the ready-to-use APK is provided.		

1.1.4 Facebook Social Aggregator

Directory	Software/public/TUD_SocialAggregator.zip		
Collected Dataset(s)	None		
Software Description	The Social Aggregator offers three functions: 1. Crawler: A crawler has been implemented which crawls all posts from the users feed after he grants the crawler access. The data is stored in a data base. 2. Social Graph: A visualization of the user and her friends are shown in a circle where the user resides at the centre. From the user to each friend, an edge is drawn which thickness is adapted to the amount of posts the user got from this friend. Each edge is labelled with the absolute number of posts counted for this friend, regarding the user's whole Facebook history. All edges are coloured green, except these for the top ten friends, based on their posts visible to the user, which are red coloured. The edges are sorted and each friend as well as the user (all nodes) is represented by a small version of their Facebook picture. 3. Feed: Out of the data retrieved by the crawler a new feed similar to the one Facebook offers is provided. This feed can currently re-ordered based on three criteria's: number of likes, number of comments and by date.		
Measured Parameters	Facebook posts, their related comments, number of likes and comments, type of the post. The re-organized feed is only able to show pictures and videos, currently.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java, JavaScript	**Supported Operating Systems**	Tested for Firefox and Chrome under Windows 7 and Ubuntu 13.10
Software License	The software might be freely used and shared as it is.	**Dissemination Level**	Public

Publication Type	Open Source	**Requirement to obtain the software**	None
Official URL	-		
Papers Published	-		
Software Installation	Instructions how to setup the tool and how to start the different functionalities are described in depth by the readme file included in the software package.		

1.2 Twitter

1.2.1 Twitter Locality

Directory	Software/public/UC3M_TwitterLocalityCrawler.zip		
Collected Dataset(s)	-		
Software Description	Twitter Locality software makes use of Sue Moon dataset (http://an.kaist.ac.kr/traces/WWW2010.html) and starts crawling from those users getting its geographic location and followers information.		
Measured Parameters	For each user: Number of followers, list of followers, list of friends and coordinates of its main location.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	All – Tested in Ubuntu 12.04
Software License	GPL	**Dissemination Level**	Public
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities
Official URL	http://acaro.it.uc3m.es/socialTools/		
Papers Published	R. Cuevas, R. Gonzalez, A. Cuevas, C. Guerrero. “Understanding the locality effect in Twitter: measurement and analysis.” Personal and Ubiquitous Computing, 2014.		
Software Installation	In order to avoid Twitter limitation, a cluster of machines could be needed.		

1.2.2 Twitter Crawler (Tweet + Trends)

Directory	Software/public/UC3M_TwitterTrending.zip

Collected Dataset(s)	-		
Software Description	Twitter Crawler is divided in: Tweet crawler which gets all available tweets using Streaming API from an input word. Trend crawler which continuously obtain the top 10 trending topics from a particular location. A usage example is to get those tweets related with a Trending Topic. Twitter API: REST API and Streaming API		
Measured Parameters	List of trending topics and tweets associated.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	All – Tested in Ubuntu 12.04
Software License	GPL	**Dissemination Level**	Public
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities
Official URL	http://acaro.it.uc3m.es/socialTools/		
Papers Published	J. Carrascosa, R. Gonzalez, R. Cuevas, A. Azcorra: "Are Trending Topics useful for marketing? Visibility of Trending Topics vs Traditional Advertisement". ACM Conference on Online Social Networks (COSN 2013).		
Software Installation	In order to avoid Twitter limitation, a cluster of machines could be needed. Limitations: - REST API: 150 query/hour per IP. - Streaming API: Best effort service		

1.3 Google+

1.3.1 Google+ Crawler

Directory	Software/public/UC3M_GplusActivityCrawler.zip
Collected Dataset(s)	G+ Connectivity and Profile Data
Software Description	The software is composed of a web crawler and several tools which use the G+ API. The web crawler starts a Binary Search Function (BSF) crawling from an initial list of users (it can start even with a single user) and capture all the users (profile and connectivity data) reachable for the firsts ones in less than 15 days.

	The API crawlers collects the activity information of the users from a list of user Ids.		
Measured Parameters	Profile and connectivity information. Users public activities.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	Only tested in Linux, but is Java, thus, it should be multi-platform.
Software License	GPL	**Dissemination Level**	Public
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities
Official URL	http://acaro.it.uc3m.es/socialTools/		
Papers Published	Roberto Gonzalez, Rubén Cuevas, Reza Motamedi, Reza Rejaie and Angel Cuevas: Google+ or Google-?: Dissecting the evolution of the new OSN in its first year (Accepted for publication in WWW'13)		
Software Installation	In order to avoid G+ limitations, a cluster of machines could be needed.		

1.4 YouTube

1.4.1 YouTube Crawler

Directory	Software/public/IMDEA-YouTubeCrawler.zip		
Collected Dataset(s)	YouTube Data		
Software Description	This crawler uses the APIs of YouTube to retrieve statistics of a video. To retrieve the statistics, the crawler has to have a valid video ID. The crawler obtains a list of recently uploaded videos and then, it retrieves information and statistics of those videos.		
Measured Parameters	Statistics of the video: number of views, comments, likes, dislikes, rating, number of raters, duration, uploader information (Number of subscribers, views uploader and country (if available)). YouTube Insight Information: top ten traffic sources to the video (Key discovery events) Other: Statistics of YouTube Insight cannot be retrieve anymore due to an update of YouTube.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative

Programming Language	PHP	**Supported Operating Systems**	All
Software License	GPL	**Dissemination Level**	Public
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities
Official URL	http://acaro.it.uc3m.es/robot		
Papers Published	-		
Software Installation	The use of this crawler requires PHP and MySQL. The instructions to execute the crawler are included in the Readme file of the software. Limitation: Due to the limitation of queries of YouTube API, the crawler is able to retrieve ~3600 videos per hour per IP.		

1.5 BitTorrent

1.5.1 BitTorrent Macroscopic Crawler

Directory	Software/public/UC3M_BTPeerRequester.zip		
Collected Dataset(s)	BT Macroscopic Data		
Software Description	The software connects to "The Pirate Bay" and monitors every new torrent published. Then the software periodically connects to the tracker in order to obtain the IP and user name of the initial seeder and the IP addresses of the downloaders. This crawler can monitor thousands of torrents simultaneously.		
Measured Parameters	Number of torrents, and the number of seeders and leechers of each torrent and their IP addresses. Also information regarding each torrent such as type of content (e.g. movies, video, audio, etc.)		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	Only tested in Linux, but is Java, thus, it should be multi-platform.
Software License	GPL	**Dissemination Level**	Public
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities

Official URL	http://acaro.it.uc3m.es/socialTools/
Papers Published	M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: “Measuring BitTorrent Ecosystem: Techniques, Tips and Tricks”, IEEE Communications Magazine, Vol. 49, Issue 9, pp. 144-152, September 2011. M. Kryczka, R. Cuevas, C. Guerrero, A. Azcorra: “Unrevealing the structure of live BitTorrent Swarms: methodology and analysis”, IEEE International Conference on Peer-to-Peer Computing P2P 2011, Kyoto, Japan, 2011 M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: “Understanding the connectivity properties of real BitTorrent swarms and their implications in swarming efficiency, resilience and locality”, under submission. R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, R. Rejaie: “Is Content Publishing in BitTorrent Altruistic or Profit Driven?”, The 6th International Conference on emerging Networking EXperiments and Technologies (CoNEXT), Philadelphia, USA, 2010 R. Cuevas, M. Kryczka, Á. Cuevas, S. Kaune, R. Rejaie, C. Guerrero:”Unveiling the Incentives for Content Publishing in Popular BitTorrent Portals” IEEE/ACM Transactions on Networking. Oct. 2013. Rubén Cuevas, Michal Kryczka, Roberto González, Angel Cuevas, Arturo Azcorra: “TorrentGuard: Stopping scam and malware distribution in the BitTorrent ecosystem”, Computer Networks, February 2014
Software Installation	In order to avoid being banned from the tracker it is required to use 6-10 different machines to obtain a complete snapshot.

1.5.2 BitTorrent Microscopic Crawler

Directory	Contact Roberto Gonzalez (rgonza1@it.uc3m.es) or Ruben Cuevas (rcuevas@it.uc3m.es)		
Collected Dataset(s)	-		
Software Description	This software periodically connects to the tracker in order to obtain the IP and user name of the initial seeder and the IP addresses of the downloaders. It also uses the Peer Exchange system (PEX) to obtain the neighbours list and the pieces already downloaded for each peer. This crawler can monitor a small number of torrents simultaneously.		
Measured Parameters	Number of torrents, and the number of seeders and leechers of each torrent. For each leecher also the neighbours list and the number of pieces already downloaded.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	Only tested in Linux, but is Java, thus, it should be multi-platform.

Software License	GPL	**Dissemination Level**	Confidential
Publication Type	Open Source	**Requirement to obtain the software**	Only available for research activities (upon acceptance of UC3M)
Official URL	http://acaro.it.uc3m.es/socialTools/		
Papers Published	M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: “Measuring BitTorrent Ecosystem: Techniques, Tips and Tricks”, IEEE Communications Magazine, Vol. 49, Issue 9, pp. 144-152, September 2011. M. Kryczka, R. Cuevas, C. Guerrero, A. Azcorra: “Unrevealing the structure of live BitTorrent Swarms: methodology and analysis”, IEEE International Conference on Peer-to-Peer Computing P2P 2011, Kyoto, Japan, 2011. M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: “Understanding the connectivity properties of real BitTorrent swarms and their implications in swarming efficiency, resilience and locality”, under submission. R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, R. Rejaie: “Is Content Publishing in BitTorrent Altruistic or Profit Driven?”, The 6th International Conference on emerging Networking EXperiments and Technologies (CoNEXT), Philadelphia, USA, 2010. M. Kryczka, R. Cuevas, R. Gonzalez, A. Cuevas, A. Azcorra: “TorrentGuard: stopping scam and malware distribution in the BitTorrent ecosystem”, under submission. R. Cuevas, M. Kryczka, Á. Cuevas, S. Kaune, R. Rejaie, C. Guerrero: ”Unveiling the Incentives for Content Publishing in Popular BitTorrent Portals” IEEE/ACM Transactions on Networking. Oct. 2013. Rubén Cuevas, Michal Kryczka, Roberto González, Angel Cuevas, Arturo Azcorra: “TorrentGuard: Stopping scam and malware distribution in the BitTorrent ecosystem”, Computer Networks, February 2014.		
Software Installation	In order to avoid being banned from the tracker it is required to use 6-10 different machines to obtain a complete snapshot		

1.6 BTLive

1.6.1 BTLive Wireshark Plugin

Directory	Software/public/TUD_BTLive-Wireshark-Plugin.zip
Collected Dataset(s)	BTLive Traces
Software Description	To study and understand the young and promising peer-to-peer live streaming protocol BTLive and its performance under real-world conditions, a BTLive Wireshark plugin was implemented at TUD.
Measured	The plugin is able to parse exchanged BTLive packets, indicating a part of the protocol fields that were derived by studying the communication patterns, the

Parameters	size characteristics of the messages, and easily to infer field types, such as IP addresses.		
Measurement Class	Passive	**Measurement Environment**	Non-cooperative
Programming Language	C	**Supported Operating Systems**	Ubuntu (amd64 and x86)
Software License	The binary might be freely used and shared as it is.	**Dissemination Level**	Public
Publication Type	Binary	**Requirement to obtain the software**	None
Official URL	http://www.ps.tu-darmstadt.de/research/btlive/		
Papers Published	-		
Software Installation	For ease of simple installation, builds for Ubuntu 32/64bit are provided.		

1.6.2 BTLive Web Crawler

Directory	Software/public/TUD_BTLive-Webcrawler-source.zip		
Collected Dataset(s)	Not yet ready for publication		
Software Description	To investigate the popularity and use of the BitTorrent Live beta version, two crawling tool were implemented to (1) derive information on new channels and (2) collect statistics for the individual channels. These two components are included in this software package.		
Measured Parameters	In particular the two components measure a number of parameters. The Google Crawler is responsible to discover new BTLive channels and to add them to the list of already known channels. The list of all channels is managed in a database and directly updated by the crawler component. The BitTorrent Web Crawler accesses the individual channel webpage and extracts a set of meta information on the. The most relevant meta information to monitor the channel activity is the channel status (ON AIR/OFF AIR) and the total number of views.		
Measurement Class	Active	**Measurement Environment**	Non-cooperative
Programming Language	Java	**Supported Operating Systems**	No limitations (Java)
Software License	The source code may be freely used as it is.	**Dissemination Level**	Public

Publication Type	Open Source	**Requirement to obtain the software**	None
Official URL	-		
Papers Published	-		
Software Installation	The component is provided as Java source files that can be run with any Java version greater or equal JRE 1.5.		

1.7 Monitoring in Operational Networks

1.7.1 Mobile Bandwidth Measurement App (NetworkCoverage)

Directory	Software/public/TUD_NetworkCoverageApp_v0.2.1(7).apk		
Collected Dataset(s)	Mobile Bandwidth Traces (NetworkCoverage)		
Software Description	Map the networks you are using and measure the actual performance. Support the creation of a network quality map to find places with the best coverage.		
Measured Parameters	Cellular: Signal strength, Arbitrary strength unit (ASU), network type, network operator, cell identifier, location area code (LAC) WiFi: Signal strength (in dB and converted to a level between 0 and 15), Service set identification (SSID), basic service set identification (BSSID) Cellular or WiFi measurements may be associated with one or both of: round-trip-time (minimum, average, maximum, mean deviation), throughput (down) Each measurement also contains: location (latitude, longitude, velocity, accuracy), time stamp		
Measurement Class	Active/passive	**Measurement Environment**	Cooperative
Programming Language	Java	**Supported Operating Systems**	Android
Software License	The binary might be freely used and shared as it is.	**Dissemination Level**	Public
Publication Type	Binary	**Requirement to obtain the software**	None
Official URL	https://play.google.com/store/apps/details?id=de.tudarmstadt.networkcoverage		
Papers Published	-		
Software Installation	Install on Android using Google Play		

1.8 Prediction Software

1.8.1 Simulator for Content Prediction

Directory	Contact Ali Gouta (ali.gouta@orange.com) or Yannick Le Louedec (yannick.lelouedec@orange.com)		
Collected Dataset(s)	Non applicable		
Software Description	The purpose of the pre-fetching simulator "Prefsim" is to assess the performance of content prediction algorithms: Try to find out which contents should be prefetched based on users' preferences and social or interest relationships.		
Measured Parameters	Hit-ratio (HR): Ratio of items that have been pre-fetched AND requested by the clients, out of the total number of requests. Correct Prediction Ratio (CPR): Ratio of items that have been pre-fetched AND requested by the clients, out of the total prefetched items.		
Measurement Class	Non applicable	**Measurement Environment**	Non applicable
Programming Language	Java	**Supported Operating Systems**	All
Software License	None	**Dissemination Level**	Restricted
Publication Type	Open Source	**Requirement to obtain the software**	None
Official URL	Non applicable		
Papers Published	-		
Software Installation	To be included in the class-path: jdom-1.1.2.jar, mahout-core-0.8-job.jar		

2. DATASETS AND TRACES

This section provides an outline of the initial datasets and traces that have been gathered with the above tools. This includes datasets of OSNs such as Facebook and Google+, as well as datasets of content distribution and content portals such as YouTube, BitTorrent, and BTLive. The section is completed with a description of traces from operational networks and with eye tracking and EEG data for modelling social cascades.

A detailed overview on the different datasets and traces with their dissemination level is given below. (Public: Available publicly; Restricted/Confidential: Available upon request only).

	Datasets / Traces	Dissemination Level
Facebook	Facebook Profiles Connectivity Data	Public
	Facebook Profiles Data	Restricted
	Facebook Brands Data	Restricted
G+ Connectivity and Profile Data		Public / Restricted
YouTube Data		Public
BT Macroscopic Data		Public
BTLive Traces		Public
Mobile Bandwidth Traces (NetworkCoverage)		Confidential
Eye Tracking and EEG Data for Modelling Social Cascades		Restricted

2.1 Facebook

2.1.1 Facebook Profiles Connectivity Data

Directory/URL	Datasets/public/TSP_Facebook-Profiles-Connectivity-Data.rar		
Collection Software	Not available		
Data Description	This dataset includes information of profile attributes for near to half million regular users includes they friend list.		
Key Figures	479K users		
Measurement Duration	Feb. 2012 to Mar. 2012		
Trace is derived	There are both raw traces as well as DB processed traces.	**Format**	Raw data in text format
Data Limitations	The dataset includes only information that is publicly available.		
Sanitization/ Anonymization	All the users' IDs are anonymized.		
Dissemination Level	Public	**Requirement to obtain the data**	Only available for research activities
Official URL	-		
Papers Published	R. Farahbakhsh, X. Han, A. Cuevas, N. Crespi, Privacy Evolution of Publicly Disclosed Information in Facebook Profiles, IEEE/ACM ASONAM 2013, Niagara Falls, Canada, August 25-28, 2013 W. Chanthaweethip, X. Han, N. Crespi, R. Farahbakhsh, A. Cuevas, "Current City" Prediction for Coarse Location Based Applications on Facebook," IEEE GLOBECOM 2013, USA, December 2013		

2.1.2 Facebook Profiles Data

Directory/URL	Contact Reza Farahbakhsh (reza.farahbakhsh@it-sudparis.eu) or Angel Cuevas Rumin (angel.cuevas_rumin@it-sudparis.eu)		
Collection Software	Not available		
Data Description	This dataset includes information of friend list attribute information (connectivity of users) for near to half million regular users.		
Key Figures	479K users		
Measurement Duration	Feb. 2012 to Mar. 2012		
Trace is	There are both raw traces as	**Format**	xml

derived	well as DB processed traces.		
Data Limitations	The dataset includes only information that is publicly available.		
Sanitization/ Anonymization	All the users' IDs are anonymized.		
Dissemination Level	Restricted	**Requirement to obtain the data**	Only available for research activities (upon acceptance of TSP)
Official URL	-		
Papers Published	R. Farahbakhsh, X. Han, A. Cuevas, N. Crespi, Privacy Evolution of Publicly Disclosed Information in Facebook Profiles, IEEE/ACM ASONAM 2013, Niagara Falls, Canada, August 25-28, 2013 W. Chanthaweethip, X. Han, N. Crespi, R. Farahbakhsh, A. Cuevas, "Current City" Prediction for Coarse Location Based Applications on Facebook," IEEE GLOBECOM 2013, USA, December 2013		

2.1.3 Facebook Brands Data

Directory/URL	Contact Reza Farahbakhsh (reza.farahbakhsh@it-sudparis.eu) or Angel Cuevas Rumin (angel.cuevas_rumin@it-sudparis.eu)		
Collection Software	Facebook Brands Crawling (Restricted)		
Data Description	This dataset includes 50 professional users activity information includes all their published posts in their wall pages.		
Key Figures	50 professional activity information		
Measurement Duration	April 2013		
Trace is derived	There are both raw traces as well as DB processed traces.	**Format**	Raw data in text format
Data Limitations	The dataset includes only information that is publicly available.		
Sanitization/ Anonymization	-		
Dissemination Level	Restricted	**Requirement to obtain the data**	Only available for research activities (upon acceptance of TSP)
Official URL	-		
Papers Published	-		

2.2 Google+

2.2.1 G+ Connectivity and Profile Data

Directory/URL	http://acaro.it.uc3m.es/socialTools/gplusGraph/		
Collection Software	Google+ Crawler		
Data Description	The dataset is composed of 14 snapshots (1 or 2 per month) of the public connectivity and profile information of the users in the largest connected component of G+.		
Key Figures	More than 190M users with more than 4B relationships among them		
Measurement Duration	April 2012 – June 2013		
Trace is derived	False	**Format**	MySQL/CSV
Data Limitations	It only captures the public attributes. If a user is added to the system between the start and the end of the capture, we probably miss it.		
Sanitization/ Anonymization	The user ids have been anonymized.		
Dissemination Level	Connectivity – Public Profile info - Restricted	**Requirement to obtain the data**	Only available for research activities (upon acceptance of UC3M)
Official URL	http://acaro.it.uc3m.es/socialTools/		
Papers Published	Roberto Gonzalez, Rubén Cuevas, Reza Motamedi, Reza Rejaie and Angel Cuevas: Google+ or Google-?: Dissecting the evolution of the new OSN in its first year (Accepted for publication in WWW'13)		

2.3 YouTube

2.3.1 YouTube Data

Directory/URL	Datasets/public/IMDEA-youtube.zip		
Collection Software	YouTube Crawler		
Data Description	The dataset is composed of two datasets of recently uploaded videos, retrieving statistics of these videos every hour.		
Key Figures	60880 videos, one measure of these videos every hour		
Measurement Duration	From 1st of February, 2013 (1st dataset) and 20th of February (2nd dataset), 2013 until 8th of April, 2013		
Trace is derived	False	**Format**	MySQL

Data Limitations	Data depends on public statistics available		
Sanitization/ Anonymization	Data does not need to be anonymized as only public statistics were obtained.		
Dissemination Level	Restricted	**Requirement to obtain the data**	Only available for research activities
Official URL	-		
Papers Published	-		

2.4 BitTorrent

2.4.1 BT Macroscopic Data

<table>
<tr><td><b>Directory/URL</b></td><td colspan="3">Datasets/public/UC3M_BTMacroscopicData.zip</td></tr>
<tr><td><b>Collection Software</b></td><td colspan="3">BitTorrent Macroscopic Crawler</td></tr>
<tr><td><b>Data Description</b></td><td colspan="3">The RSS feed provides the .torrent file along with the username of the content publisher. Our tool retrieves the IP address of the tracker from the .torrent file (or the magnet link) and immediately connects to it. By connecting to the tracker right after the content is published, we are able to identify the IP address of the initial seeder (i.e. the publisher’s location) in many torrents. Our tool periodically connects to the tracker to retrieve the IP addresses for (typically) 200 randomly-selected participating peers (i.e. consumers) while respecting the reconnection time imposed by the tracker in order to avoid being banned. To cope with this limitation, our tool probes a tracker from eight geographically- distributed nodes in parallel and captures the IP address of a majority of consumers.</td></tr>
<tr><td><b>Key Figures</b></td><td colspan="3"><table>
<tr><th></th><th>Trace2009</th><th>Trace2011</th><th>Trace2012</th></tr>
<tr><td>Publishers (username)</td><td>7.1K</td><td>6.9K</td><td>3.3K</td></tr>
<tr><td>Torrents</td><td>38.2K</td><td>72.0K</td><td>21.0K</td></tr>
<tr><td>Consumers</td><td>27.3M</td><td>25.6M</td><td>5.1M</td></tr>
<tr><td>Downloads</td><td>95.6M</td><td>79.0M</td><td>11.1M</td></tr>
</table></td></tr>
<tr><td><b>Measurement Duration</b></td><td colspan="3">Trace2009 – 28.11.2009/18.12.2009
Trace2011 – 06.04.2010/05.05.2010
Trace2012 – 30.04.2011/13.05.2011</td></tr>
<tr><td><b>Trace is derived</b></td><td>False</td><td><b>Format</b></td><td>MySQL/CSV</td></tr>
<tr><td><b>Data Limitations</b></td><td colspan="3">In torrents with a lot of simultaneous leechers we cannot have the information of some of them.</td></tr>
</table>

Sanitization/ Anonymization	The IP addresses have been anonymized.		
Dissemination Level	Public	**Requirement to obtain the data**	Only available for research activities
Official URL	http://acaro.it.uc3m.es/socialTools/btAnonymized/		
Papers Published	M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: “Measuring BitTorrent Ecosystem: Techniques, Tips and Tricks”, IEEE Communications Magazine, Vol. 49, Issue 9, pp. 144-152, September 2011. M. Kryczka, R. Cuevas, C. Guerrero, A. Azcorra: “Unrevealing the structure of live BitTorrent Swarms: methodology and analysis”, IEEE International Conference on Peer-to-Peer Computing P2P 2011, Kyoto, Japan, 2011. M. Kryczka, R. Cuevas, A. Cuevas, C. Guerrero, A. Azcorra: “Understanding the connectivity properties of real BitTorrent swarms and their implications in swarming efficiency, resilience and locality”, under submission. R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, R. Rejaie: “Is Content Publishing in BitTorrent Altruistic or Profit Driven?”, The 6th International Conference on emerging Networking EXperiments and Technologies (CoNEXT), Philadelphia, USA, 2010. R. Cuevas, M. Kryczka, Á. Cuevas, S. Kaune, R. Rejaie, C. Guerrero:”Unveiling the Incentives for Content Publishing in Popular BitTorrent Portals” IEEE/ACM Transactions on Networking. Oct. 2013. Rubén Cuevas, Michal Kryczka, Roberto González, Angel Cuevas, Arturo Azcorra: “TorrentGuard: Stopping scam and malware distribution in the BitTorrent ecosystem”, Computer Networks, February 2014.		

2.5 BTLive

2.5.1 BTLive Traces

Directory/URL	Datasets/public/TUD_BTLive-SampleTraces.zip
Collection Software	BTLive Wireshark Plugin
Data Description	The dataset includes two typical BitTorrent Live streaming sessions as captured using the BTLive Wireshark Plugin. In both sessions, first the source was started and 10 peers were subsequently added to the swarm, with an offset of 2 seconds. The .pcap files were captured on the last peer that joined the swarm. The peers were running on EmanicsLab servers (http://www.emanicslab.org) on 10 different sites across Europe. Streaming was terminated at all peers after 5 minutes. The capture length for each packet was limited to 200 bytes.
Key Figures	Involved entities in the session are: the tracker, the server, and eleven peers. The first dataset includes 312,571 captured packets, the second dataset 33,941.

Measurement Duration	First dataset: 2014-02-27 16:08-16:13 Second dataset: 2014-03-13 14:31-14:33		
Trace is derived	False	**Format**	Pcap file following the BTLive message format
Data Limitations	-		
Sanitization/ Anonymization	IP addressed might be removed to be sure no real users were captured by chance.		
Dissemination Level	Public	**Requirement to obtain the data**	-
Official URL	-		
Papers Published	-		

2.6 Monitoring in Operational Networks

2.6.1 Mobile Bandwidth Traces (NetworkCoverage)

Directory/URL	Contact Fabian Kaup (fkaup@ps.tu-darmstadt.de) or David Hausheer (hausheer@ps.tu-darmstadt.de)		
Collection Software	Mobile Bandwidth Measurement App (NetworkCoverage)		
Data Description	Measurements of the cellular network availability and QoS, currently focused on the Darmstadt (Germany) region.		
Key Figures	• 200k cell coverage measurements • 22k RTT measurements (18k cell) • 800 downlink measurements (500 cell)		
Measurement Duration	Continuous since 2013-10-25		
Trace is derived	False	**Format**	Postgres/PostGis DB
Data Limitations	RTT measurements miss the loss rate. Network transitions can be filtered out.		
Sanitization/ Anonymization	No personally identifiable data is collected.		
Dissemination Level	Confidential	**Requirement to obtain the data**	Upon request only
Official URL	-		

Papers Published	-

2.7 Social Cascade Prediction

2.7.1 Eye Tracking and EEG Data for Modelling Social Cascades

Directory/URL	Contact Eiko Yoneki (Eiko.yoneki@cl.cam.ac.uk)		
Collection Software	Matlab based programming using EEG experimental tool		
Data Description	This eye tracking and EEG data has been collected in a few experiments to learn how social influences appear when the experiment's participant is making a decision on photo rating. This is part of understanding the information diffusion process, i.e. to predict propagation rate considering the social influence. This data can be used for modelling the social cascade, which helps understanding a heavy tail part of content access in the Internet.		
Key Figures	EEG data measured along media based decision making experiment		
Measurement Duration	Summer 2013		
Trace is derived	No	**Format**	Matlab data (see readme in Zipped file)
Data Limitations	5 social cliques consisting of 4 members – as EEG data, it is reasonable experiment size. Ideally same experiment using Twitter chat could reveal further interesting results.		
Sanitization/ Anonymization	Current data is not anonymised and it requires careful treat on processing data.		
Dissemination Level	Restricted	**Requirement to obtain the data**	Contact to Eiko.yoneki@cl.cam.ac.uk
Offical URL	No official URL and all information is kept in SVN maintained by the computer laboratory.		
Papers Published	-		