WSDM 2013 Data Challenge

Warning: This document is to be considered a draft until August 31, 2012. After that date, the final version will be realeased, along with all the data needed for the competition.

Contents

1 Dataset description

The data used within the challenge are provided by Microsoft Corporation and come from their Microsoft Academic Search database. Microsoft Academic Search is a free academic search engine that was developed by Microsoft Research, and covers more than 38 million publications and over 19 million authors across a variety of domains, with updates added each week.

In particular, the two tasks of the challenge deal with two graphs that are extracted from a snapshot (of May 18, 2012) of the Microsoft Academic Search database:

- The citation graph C: a directed graph whose nodes represent publications, having a directed arc from a node x to a node y if and only if the publication x contains y among its references.
- The coauthorship graph A: an undirected graph whose nodes represent authors, with an undirected arc connecting two authors x and y if and only if x and y are co-authors of some publication.

Note that both graphs have natural labels on their nodes: in the case of C the labels are publication titles; in the case of A the labels are author names. The two tasks of this challenge use these graphs and suitable subgraphs thereof, as explained in the tasks' description.

1.1 Graph data format

All the graphs that are provided are stored in text files with the extension `.graph.txt`. Any such file has the following format[1]:

n	
x_1	y_1
x_2	y_2
$\cdots$	
x_m	y_m

here n is the number of nodes of the graph, m is the number of (directed or undirected) arcs, and each pair (x_i, y_i) corresponds to an arc going from x_i to y_i (in the case of undirected graphs: an arc connecting x_i to y_i, with the proviso that $x_i < y_i$). *Nodes are numbered from 0 to* $n - 1$.

Labels for the nodes of a graph, when provided, are stored in a file with the same base name as the graph but with the extension `.names`. Such a file has the following format:

x_0
x_1
$\cdots$
x_{n-1}

where x_i is the label of node i. The file is written using the UTF-8 encoding.

[1] Values appearing in the same line are TAB-separated. Each line ends with a newline character, in the standard UNIX format.

2 Participating in the challenge

The procedure to participate in the challenge is as follows:

- The participating team has to notify the data challenge committee who is the leader of the team and which of the two tasks they are going to undertake (possibly both), by sending an e-mail to `data-challenge@wsdm2013.org` no later than September 30, 2012.
- The team leader will be required to sign and fax the Non-Disclosure Agreement provided in the Appendix of this document.
- After receiving the signed agreement, the team will be allowed to download the datasets needed for solving the chosen tasks.
- The material to be submitted (detailed in the description of each task) should be sent to `data-challenge@wsdm2013.org` by December 14, 2012.
- At least one author of each of the submissions to the data challenge must register to the WSDM conference prior to December, 14th. Submissions for which none of the authors will be registered by that date will not be evaluated.
- The registered participants will present a poster describing their work in a special poster session during the conference.

2.1 Data Challenge Committee

- Co-Chairs:
  - Paolo Boldi (Università degli Studi di Milano)
  - Tamir Tassa (The Open University of Israel)
- Committee:
  - Graham Cormode (AT&T Labs)
  - Ravi Kumar (Google)
  - Theodoros Lappas (Boston University)
  - Marc Najork (Microsoft)
  - Massimo Santini (Università degli Studi di Milano)
  - Evimaria Terzi (Boston University)

2.2 Dates

- Launching the data challenge: Aug 31, 2012
- End of data challenge: Dec 14, 2012
- Data challenge presentations: Feb 5, 2013

2.3 Acknowledgements

The co-chairs of the Data Challenge Committee would like to thank Evimaria Terzi, Aristides Gionis, and Dror Cohen for their help in producing the anonymized graphs.

3 Graph compression task

The purpose of this task is to provide a compressed data structure for storing the graphs; the data structure should allow access without full decompression. This task is described below as applied to directed graphs; this description holds also for undirected graphs, provided that an undirected graph is seen as a directed symmetric graph (i.e., an undirected arc that connects the nodes x and y is represented by the pair of directed arcs (x, y) and (y, x)).

This task requires providing, for a directed graph G with n nodes, a compact data structure with the following properties:

- The nodes of G are represented internally in some manner by *node handlers* (that, in the most trivial case, are just integers from 0 to $n-1$; note that in such a case it is not required that the order of the nodes is the same as the one in which the dataset is provided).
- There is a way to map the node handlers used by the data structure to their original number, and vice-versa; how this bijection is provided, its efficiency or memory footprint are not considered in the evaluation of the task.
- The time needed to build the data structure is not part of the evaluation, either, provided that it is reasonable.
- The data structure should allow for some or all of the following access modes:
  - *sequential acces*: all node handlers are provided in some sequence, and after each node handler x all the out-neighbors y of x are listed;
  - *sequential bidirectional access*: all node handlers are provided in some sequence, and after each node handler x all the out-neighbors y and all in-neighbors z of x are separately listed;
  - *direct access*: given a node handler, all its out-neighbors are listed;
  - *direct bidirectional access*: given a node handler, all its out-neighbors and all its in-neighbors are listed, separately;
  - *boolean access*: given two node handlers x and y, it is possible to know whether (x, y) is an arc or not.

The average time required for each of these access modes is part of the evaluation; that average time will be computed for all the contestants using the same sequence of accesses. The data structure should be as memory-efficient as possible; the memory efficiency is evaluated looking at the actual average-case quantity of main memory occupied by the data structure while accessing the graph. It is probably going to be different depending on the access mode; also, different forms of memory usage can be offered, providing different space/time trade-offs. The memory footprint is evaluated in bits/arc.

Participants are supposed to provide all the details about the hardware they used to run the code.

3.1 Datasets provided for this task

For this task, we will provide three groups of graphs: *lightweight* (with no more than 100 000 nodes), *middleweight* (with more than 100 000 but less than 1 000 000 nodes),

and *heavyweight* (with more than 1 000 000 nodes). For each group, we will provide some induced subgraphs of the citation graph (that are genuinely directed) and some subgraphs of the coauthorship graph (that, as explained above, should be viewed as symmetric directed graphs).

3.2 What are participants supposed to submit

Participants are supposed to submit:

- A report explaining the compression technique they exploited. The report should be written in the WSDM paper format and is limited to four pages.
- The results (compression ratio, compression time, access time etc.) they obtained on the datasets provided, with a detailed description of the hadware they used. Note that the applied techinque may work only for some of the datasets: for example, it is legitimate to submit an algorithm that only works for undirected graphs, or that could only be applied to middleweight and lightweight graphs, but not to larger ones: the authors should state explicitly on which datasets they could run their code, and why.
- The compressed graphs obtained.
- The software needed to access the compressed graphs, with full instructions needed to install and run it on a Linux machine, and code snippets showing how to actually access the graphs.

4 De-anonymization task

An *anonymized* version of a graph is another graph that supposedly resembles the original one (as far as most data mining features are concerned) but makes it difficult to identify specific nodes based on the knowledge of some of their properties (e.g., their degree in the original graph, or the structure of their neighborhood).

For this task, we start by considering the coauthorship graph A. Given a set X of authors, we let $A[X]$ denote the subgraph of A induced by X (that is, the graph whose nodes are the elements of X and having an undirected arc connecting two of them iff there is a corresponding arc in A).

We selected as X the set of authors who published a paper in one of the leading data mining conferences since 2007. The list X will be disclosed to the contestants. Namely, the contestants will have the graph $A[X]$ together with the labels that associate with each node the name of the corresponding author.

4.1 Datasets provided for this task

For this task, we shall provide several anonymized versions of $A[X]$, which we denote by $H_1, \ldots, H_t$. The anonymizations will be created by some of the methods that were published in recent years for graph anonymization:

1. k-degree anonymity (Kun Liu, Evimaria Terzi: Towards identity anonymization on graphs. SIGMOD 2008: 93-106; downloadable from `http://www.csd.uoc.gr/~hy558/papers/graph_anonymity.pdf`);
2. anonymization by random perturbations (Francesco Bonchi, Aristides Gionis, Tamir Tassa: Identity obfuscation in graphs through the information theoretic lens. ICDE 2011: 924-935; downloadable form `http://www.openu.ac.il/Personal_sites/tamirtassa/Download/Conferences/granon.pdf`);
3. anonymization by clustering (Tamir Tassa and Dror Cohen: Anonymization of centralized and distributed social networks by sequential clustering, TKDE 2012; downloadable from `http://www.openu.ac.il/Personal_sites/tamirtassa/Download/Journals/cdsn.pdf`).

The selected methods represent three main paradigms in graph anonymization.

In the first two methods, the anonymized graph is a graph with the same number of nodes as A (that is, $|X|$), but its nodes are abstractly identified with numbers (from 0 to $|X| - 1$). The contestants will have to recover the linkage between nodes in each of the anonymized graphs H_i and the author names.

In the third method, the original graph $A[X]$ is released in a clustered form, where the size of each cluster is at least k, for some security threshold parameter k. The graph information is then released with respect to that clustering: each cluster is accompanied by two numbers that indicate the number of nodes in the cluster and the number of internal arcs (i.e., the number of arcs between those nodes in $A[X]$); each super-arc between two clusters is labeled by the number of original arcs in $A[X]$ that connect nodes in those two clusters. The contestants will have to disclose the links between author names and clusters in H_i.

For this task, we shall therefore provide

- the subgraph $A[X]$ (in the usual format, as an undirected graph; a label file will also be provided, where the labels represent the author names);

- anonymized versions of $A[X]$ obtained by applying methods (1) and (2); since the outputs of those methods are themselves undirected graphs (with $|X|$ nodes), they will be represented using the usual format;
- anonymized versions of $A[X]$ obtained by applying method (3); the formats for these files will be as follows:

c		
n_0	m_0	
n_1	m_1	
$\cdots$		
n_{c-1}	m_{c-1}	
x_1	y_1	e_1
$\cdots$		
x_p	y_p	e_p

  here, c is the number of clusters (numbered from 0 to $c-1$); n_i is the number of nodes in cluster i while m_i is the number of undirected arcs connecting nodes of cluster i; each of the final p lines correspond to a super-arc connecting two clusters: x_i and y_i are the numbers of the two clusters connected by the super-arc (where $x_i < y_i$), and e_i is the number of arcs of the original graph that connect a node from the first cluster to a node in the second cluster.

For method (1) (k-degree anonymity) we shall provide four anonymized versions, corresponding to four different selections of the security parameter, $k = 10, 20, 30, 40$. Similarly, also for method (3) (anonymization by clustering) we shall provide four anonymized versions that correspond to $k = 10, 20, 30, 40$, where here k denotes the lower bound on the size of clusters. As for method (2), it has two variants: one in which arcs can be only removed (obfuscation by sparsification) and another in which arcs can be also added (obfuscation by perturbation). We shall provide four versions for each of those variants, which were obtained by different settings of the level of noise applied to the graph structure.

4.2 What are participants supposed to submit

A *candidate disclosure* is a triple (i, x, v) where $i \in \{1, \ldots, t\}$, x is a node of H_i, and v is a node number in the original graph $A[X]$ (i.e., a number between 0 and $|X| - 1$, corresponding to an element of X, namely to an author). A candidate disclosure is *correct* if and only if:

- The node x in H_i is indeed the image of the node which represents author v, in the case of the first two methods;
- The node x in H_i is a cluster that includes author v, in the case of the third method.

Participants are supposed to submit a sequence of candidate disclosures that will be evaluated using the MAP metrics which is commonly used in information retrieval. More precisely, for every $i \in \{1, \ldots, t\}$, let $(i, x_1, v_1), \ldots, (i, x_s, v_s)$ be the list of candidate disclosures submitted for the i-th anonymized graph H_i (in the order in which

they were submitted), and let $r_1, \ldots, r_s$ be a sequence of 0s or 1s, where $r_j = 1$ iff the corresponding disclosure is correct. Define:

$$A_i = \sum_{j=1}^{s} \frac{r_j \sum_{k=1}^{j} r_k}{j}.$$

Then, the MAP (Mean Average Precision) is the average value of A_i over all indices $i \in \{1, \ldots, t\}$.

Participants are supposed to submit:

- A list of candidate disclosures (as explained above).
- A report explaining the technique that they exploited in order to obtain the submitted candidate disclosures. In the report they should in particular emphasize details about the computational resources that were required by their algorithm for obtaining the results. The report should be written in the WSDM paper format and is limited to four pages.

5 Appendix — Terms of use

I hereby certify that I have read and understood the Terms of Use below, and that I accept them.

Name	
Title	
Address	
Date	

Signature .....

WSDM DATA CHALLENGE MICROSOFT ACADEMIC SEARCH DATA — TERMS OF USE

THIS IS AN AGREEMENT BETWEEN YOU AND MICROSOFT

These Terms of Use ("TOU") are an agreement between you and Microsoft Corporation ("Microsoft" or "we"). This TOU governs your use of the datasets provided for the WSDM 2012 Data Challenge ("Data"); such datasets contain data extracted from Microsoft Academic Search to use in your own application. This TOU does not govern your use of other Microsoft products, services or web sites. You represent that you are at least 18 years old. Your use of the Data constitutes your acceptance of this TOU, the Privacy Statement and registration guidelines, without modification.

Please note that we do not provide warranties for the Data. The TOU also limits our liability. These terms are in the "Disclaimer of Liability and Warranty" section below, and we ask you to read it carefully.

PERSONAL USE LIMITATION

You may use these Data for any non-revenue/no-fee academic purpose, subject to the restrictions in this TOU. Some purposes which can be non-revenue/no-fee academic purpose are teaching, academic research, public demonstrations and personal experimentation. You may leverage the Data to build your own application or service. You may not use these Data or any derivative works in any form for commercial purposes. Examples of commercial purposes would be running business operations, licensing, leasing, or selling content obtained from the Data, distributing the Data for use with

commercial products, using the Data in the creation or use of commercial products or any other activity which purpose is to procure a commercial gain to you or others.

YOUR RESPONSIBILITIES

You, and any third party working on behalf of you, will not:

1. Generate revenue from use of the Data or use it for any commercial or non-academic purposes;
2. Bypass Microsoft's tools or services to interfere or attempt to interfere with the proper working of Microsoft Academic Search or its API;
3. Take any action that imposes a disproportionately large burden on use of the Data or Microsoft's infrastructure, including using the Microsoft Academic Search API on more than a reasonable frequency, as determined by Microsoft in its sole discretion; or
4. Engage in any unlawful practices in connection with your use of the Data.

TERMINATION

1. Termination by Microsoft. Microsoft may terminate or suspend this Agreement at any time in its sole discretion.
2. Termination by you. You may stop using the Data provided if you are dissatisfied with any aspect of those Data.
3. Survival. The sections entitled Your responsibilities, Termination, Disclaimer of warranties, Indemnity, Limitations of liability, Reservation of rights and Miscellaneous will survive expiration or termination of this Agreement.

DISCLAIMER OF WARRANTIES

You use the Data and any related products at your own risk. The Data and any related products are provided *as is* and with all defects. Except as provided herein, Microsoft expressly disclaims all other express, implied, or statutory warranties. This includes the warranties of merchantability, fitness for a particular purpose, title, non-infringement, lack of viruses, quiet enjoyment, scope of license, lack of errors, and satisfactory condition or quality.

INDEMNITY

You will indemnify and hold Microsoft and its affiliates, agents and employees, harmless from all loss, liability, and expense (including reasonable attorneys' fees) from any claims, proceedings or suits due to either your breach of this Agreement or your use of the Data. You will be solely responsible for defending all such claims but Microsoft may participate with counsel it selects. If Microsoft participates with its own counsel, you will reimburse Microsoft for all reasonable costs and fees incurred. You will not agree to any settlement imposing any obligation or liability on Microsoft without Microsoft's prior written consent.

LIMITATION OF LIABILITY

TO THE MAXIMUM EXTENT PERMITTED BY LAW, IN NO EVENT WILL EITHER PARTY BE LIABLE FOR ANY INDIRECT, INCIDENTAL, CONSEQUENTIAL, PUNITIVE, SPECIAL, OR EXEMPLARY DAMAGES ARISING OUT OF OR THAT RELATE IN ANY WAY TO THIS AGREEMENT OR ITS PERFORMANCE. THIS EXCLUSION WILL APPLY REGARDLESS OF THE LEGAL THEORY UPON WHICH ANY CLAIM FOR SUCH DAMAGES IS BASED, WHETHER THE PARTIES HAD BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES, WHETHER SUCH DAMAGES WERE REASONABLY FORESEEABLE, OR WHETHER APPLICATION OF THE EXCLUSION CAUSES ANY REMEDY TO FAIL OF ITS ESSENTIAL PURPOSE. THIS EXCLUSION WILL NOT APPLY TO EITHER PARTY'S LIABILITY FOR BREACH OF ITS CONFIDENTIALITY OBLIGATIONS OR VIOLATION OF THE OTHER PARTY'S INTELLECTUAL PROPERTY RIGHTS. MICROSOFT WILL NOT BE LIABLE TO YOU FOR DAMAGES IN EXCESS OF US DOLLARS 500.

RESERVATION OF RIGHTS

Microsoft reserves all rights in the Data and associated services and technologies, subject to the licenses granted in this Agreement.

MISCELLANEOUS

1. Notices. Notices may be served either by posting on any portion of the Microsoft Academic Search website or physical mail. The person who signed up for the Data Credentials will receive notices on behalf of their respective company. Each party may change the persons to whom notices will be sent by giving notice to the other. You may inform the change contact information via e-mail to acadapi@microsoft.com.

2. Jurisdiction and governing Law. The laws of the State of Washington govern this Agreement. If federal jurisdiction exists, the parties consent to exclusive jurisdiction and venue in the federal courts in King County, Washington. If not, the parties consent to exclusive jurisdiction and venue in the Superior Court of King County, Washington.

3. Attorneys' fees. If either Microsoft or you employ attorneys to enforce any rights arising out of or relating to this Agreement, the prevailing party will be entitled to recover its reasonable attorneys' fees, costs, and other expenses, including the costs and fees incurred on appeal or in a bankruptcy or similar action.

4. Assignment. You will not assign, sublicense, or otherwise dispose of this Agreement or your right to access the Data without Microsoft's prior written approval.

5. Waiver. A party's delay or failure to exercise any right or remedy will not result in a waiver of that or any other right or remedy.

6. Severability. If any court of competent jurisdiction determines that any provision of this Agreement is illegal, invalid or unenforceable, the remaining provisions will remain in full force and effect.

7. Integration and modification.

(a) Entire agreement. This Agreement (including any exhibits) is the entire Agreement between the parties regarding its subject matter. It replaces all prior agreements, communications and representations regarding its subject matter.

(b) Amendment. Microsoft may change this Agreement at any time by providing notice to you as provided in the section entitled Notices. Your continued use of the Data after receiving notice of any change will constitute your acceptance of the change.

COPYRIGHT NOTICE; TRADEMARKS