Harvesting Ego-Network Data from Facebook

Using the CEMAP Facebook Profile in ORA

Terrill L. Frantz and Kathleen M. Carley

February 2, 2009
CMU-ISR-09-102

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Center for the Computational Analysis of Social and Organizational Systems
CASOS technical report.

This work is part of the Dynamics Networks project at the center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). This work is supported in part by the Office of Naval Research (ONR), United States Navy, N00014-06-1-0104. Additional support was provided by National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) program, NSF 045 2598, the Air Force Office of Scientific Research, FA9550-05-1-0388 under a MURI on Computational Modeling of Cultural Dimensions in Adversary Organizations, the Army Research Institute W91WAW07C0063, the Army Research Lab DAAD19-01-2-0009, the Army Research Office W911NF-07-1-0060, and CASOS. The views and proposal contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Air Force Office of Scientific Research, the Army Research Institute, the Army Research Lab, the Army Research Office, the National Science Foundation, or the U.S. government.

Keywords: Facebook, CEMAP, social network, ORA, dynamic network analysis

Abstract

The Facebook social networking site (www.facebook.com) has become a popular phenomenon over the past five years. By its nature, Facebook has attracted the attention of social network researchers and enthusiasts as a rich source of data for scholarly study or personal self-analysis. It is fitting that the CEMAP software-tool, hosted by the CASOS software suite, has functionality to aid in the harvesting of Facebook data into the CASOS software, in particular, into the ORA social network analysis tool. CEMAP is designed to interface with real-world data platforms and data formats in an abstract manner that greatly reduces the need for ORA users to understand and deal with the plethora of complex technologies and data-file formats. This report introduces the CEMAP Facebook profile and provides instruction on how to use the profile—so as to easily load an individual's Facebook ego-network into ORA. We provide step-by-step instructions and illustrations as well as a description on the technology underlying the Facebook tableset. The Facebook tableset is the CEMAP abstraction of the various levels of technology to harvest the social network data, via the Facebook developer platform.

Table of Contents

1 Introduction

The Facebook social networking site (www.facebook.com) has become a popular phenomenon over the past five years. The site boosts millions of users from around the globe. By its nature, Facebook has attracted the attention of social network researchers and enthusiasts as a rich source of data for scholarly study or personal self-analysis. Unfortunately, to the average Facebook user, the ability to access their Facebook-based network of connections in a manner that allows for analysis, either visual or statistical, is out of their reach. Currently, Facebook does not provide an avenue for the non-technology-savvy user to harvest their personal network data from Facebook and into social network analysis tools. Facebook does, however, have an application called *friendwheel* that does provide a simplistic circle-formatted visualization of their network. Unfortunately, this simple application, which is embedded in the Facebook web platform, does not provide even the newest social network analysis with much flexibility or power; it certainly does not facilitate any statistical analysis of the network, whatsoever.

Fortunately, Facebook does make available a software developers platform that allows for the technology-savvy to access data via an API. The set if APIS greatly limit the extent of the data that is available through software, though it does allow for accessing the ego-network of the user. The API allows software, with the user's password-protected permission, to construct a listing of the user's friends, and the friends' friends. This essentially, formulates the user's ego-network at a distance of two. Figure 1 illustrates what a Facebook ego-network would look like graphically (this graphic is constructed using this CEMAP Facebook profile and ORA described herein).

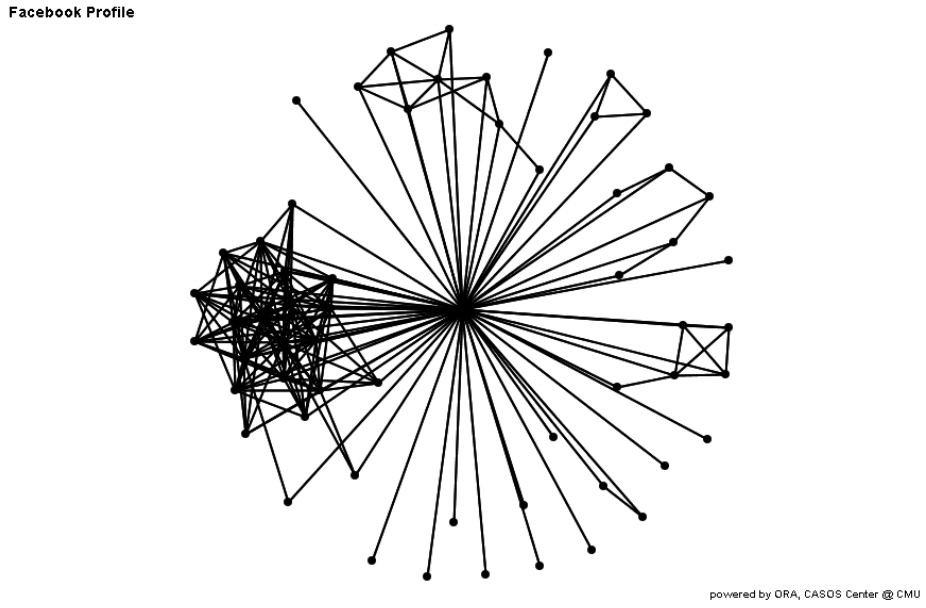


Figure 1. A ego-network from Facebook (Node names have been removed for publication purposes)

The CEMAP Facebook profile makes the process of obtaining the data from Facebook, using APIs making up the Facebook developers platform, a simple 1-2-3 process. The step-by-step instructions are provided in Section 2 of this document. CEMAP is a feature of the CASOS suite of software tools that greatly simplifies the harvesting of data from real-world sources into the CASOS software suite, which consists of a network analysis tool (ORA), a text information extraction tool (AutoMap) and a computer simulation tool (Construct). CEMAP

It is fitting that the CEMAP software-tool, hosted by the CASOS software suite, has functionality to aid in the harvesting of Facebook data into the CASOS software, in particular, into the ORA social network analysis tool. CEMAP is designed to interface with real-world data platforms and data formats in an abstract manner that greatly reduces the need for ORA users to understand and deal with the plethora of complex technologies and data-file formats. This report introduces the CEMAP Facebook profile and provides instruction on how to use the profile—so as to easily load an individual's Facebook ego-network into ORA. We provide step-by-step instructions and illustrations as well as a description on the technology underlying the Facebook tableset. The Facebook tableset is the CEMAP abstraction of the various levels of technology to harvest the social network data, via the Facebook developer platform.

1.1 CEMAP

One can easily foresee that much of a social network analyst's time is spent on messaging their raw data to fit nework software requirements, which leaves little precious time for the true value-add of the analysis step. By themselves, the analytic tools developed at the Center for Computational Analysis of Social and Organizational Systems (CASOS) do not differ from this data processing requirement. The tools included in the CASOS suite (Carley, Diesner, Reminga, & Tsvetovat, 2004) consist of the Organizational Risk Analyzer (ORA; Carley & Reminga, 2004), AutoMap (Diesner & Carley, 2004), and Construct (Carley, 1991). CEMAP addressed and provides a solution for this overbearing real-world data processing problem specifically for users of the CASOS suite of tools, by describing a simply, yet powerful, process for bridging the two separate worlds of data and analytic software.

CEMAP II's purpose is to transform real-world data into relational network data that can be used by ORA. ORA operates on data in the DyNetML format. DyNetML (Tsvetovat, Reminga, & Carley, 2003) is the XML-based document markup language that ORA relies upon. It is this format that allows for fully representation of the meta-network data and serves as the native input format of ORA. Producing network data in this DyNetML format is a primary purpose of CEMAP II. An extension to DyNetML for ORA's LOOM feature (Davis, Olsen, & Carley, 2008), call walksets is also an output of CEMAP II. CEMAP II will be expandable so as DyNetML changes, so can CEMAP II.

Is it not the purpose of this report to fully articulate CEMAP and describe its features. A full description of its purpose, architecture, and implementation is provided elsewhere (see Frantz & Carley, 2008), as is a report on details of its technology (Frantz & Carley, 2009) and use (DeReno, Frantz & Carley, 2009). See these other documents for much more detail about CEMAP.

CEMAP is available as a menu item in both ORA and AutoMap, but can also be executed independently as a stand-alone interactive program, or as a stand-alone, scripted

batch program. The primary purpose of CEMAP is to prepare and format raw source-data for use as relational network input data files for ORA, or unstructured text input files for AutoMap; although CEMAP can also produce formatted output data files for other purposes, e.g. input to Excel, etc. CEMAP's interactive interface provides a simplistic process for analysts to prepare their source data in either, a stable and routine manner, or in a frequently changing, exploratory manner. With ease, the user can repeatedly process the source data while tweaking the resulting output to their specifications. Once the CEMAP II process is producing the desired results, the profile for the process can be saved for easy re-use at a later time, by an individual or a specific group of users.

The key to CEMAP's effectiveness is the method in which it separates the physical characteristics of the source data, which is often the domain of computer programmers, from the characteristics of the stylized, reformatted meta-network and unstructured data necessary for analysis, either in ORA or AutoMap. These two very different perspectives on data are joined together, as is necessary, by a simple drag-and-drop mapping process that conjoins the source data which is viewed as straightforward, row-column structured tables, and the output data format – usually DyNetML or text files.

The cornerstone of CEMAP is the process by which the user indicates the source of the data that is transformed into the format specified by a template. This data is made available to CEMAP via a table defined in a tableset. The process by which the output template is associated with the tableset table is call "mapping." The mapping process is simplistic with very few, but strict, rules. The user will have a flexible drag and drop mechanism in the CEMAP GUI to carry out the mapping process. To set the mappings for a profile, the user will identify the template field that they are working on – the output field in the DyNetML or unstructured text file. In the case of DyNetML file, this output field could be a node, a node attribute, a link, or a LOOM walkset, among other common possibilities. Next the user indicates the specific table that contains the data that should be mapped to the output field. Once the table is identified, a specific column is selected and it is this exact column, table, and tableset combination that is associated with the output field. Once mapped, a complete profile has been created, which can be stored for later use, or executed immediately.

1.2 Facebook.com

Facebook is a very popular social networking website that came to bear in February 2004, and now just five years later its membership is up to 42 million users (iStrategyLabs, 2009). Details about Facebook is well-known and available on their website Facebook (2009) and Wikipedia (2009); instead of repeating information, that is still evolving rapidly, we suggest visiting these sites for more background information on Facebook social networking and its features.

2 Using the CEMAP Facebook Profile

This section provides step-by-step instructions on how to use the CEMAP Facebook profile in ORA to harvest your Facebook data. There are two steps: (a) authenticating the user with Facebook, and (b) performing the CEMAP process.

2.1 Authenticating with Facebook

Facebook only allows access to data to users logged into the Facebook.com site, this requirement is both for an individual's authentication and authorization. Accessing Facebook data through channels different from using a browser to connect the Facebook.com website is no different; the user needs to be identified and authorization granted. In order for the CEMAP software to access your Facebook account and portions your friends' information, you must first log into Facebook through accessing the Facebook.com website within an Internet browser.

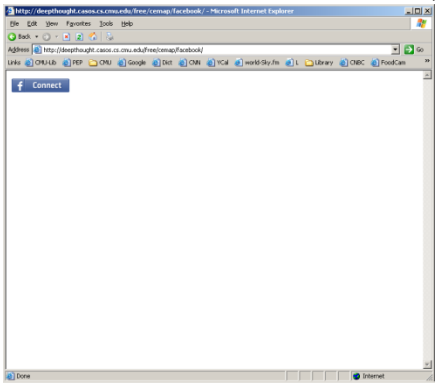
To facilitate non-Facebook applications, like CEMAP, access to Facebook data without putting the users' Facebook passwords at risk, there is a process that a user must go through to obtain a special, and temporary, access key string that the software application must then present to Facebook, via an API, to access the user's data. The application software that accesses the user's data can only have access to the data while the user is actually logged into Facebook. The special key that the user obtains, expires when the user logs out of the Facebook.com session. (note: the user need not keep Facebook.com in a browser, they can close their browser windows, without logging out of Facebook.com)

Before using CEMAP to access Facebook data, the following process must be followed. This procedure need only to be followed once during a users' Facebook.com login session. If for any reason the user logouts, or becomes logged out, of Facebook.com, these steps will need to be performed again in order to obtain a new access key string. These steps result in the user being provided a special key string that must be presented to CEMAP in order for the CEMAP software to access the data in your Facebook account.

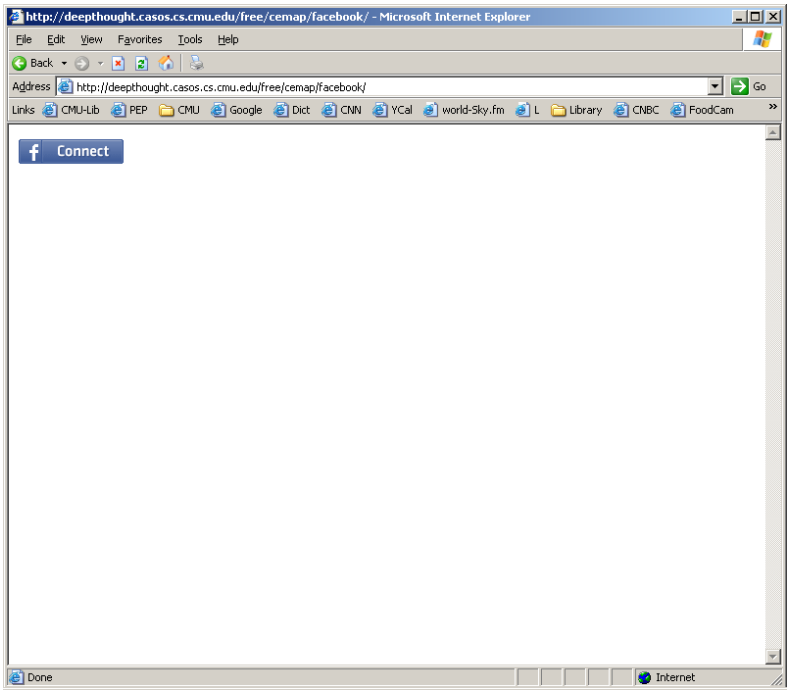
STEP 1: Open a browser window and go to URL:

http://deepthought.casos.cs.cmu.edu/free/cemap/facebook/

The page shown below will appear. (Note: this URL can be subject to change. The URL will always be in the CEMAP Facebook Profile notes (see Step 4 in the next section.)



STEP 2: Left-click the blue "Connect" button.



If you are NOT already logged into Facebook, a browser window will open asking you to log in to Facebook.com and give permission for CEMAP to access your data. If you are already logged in to Facebook, you will not see this page, instead you will see the page that follows: you will NOT need to perform STEP 2.

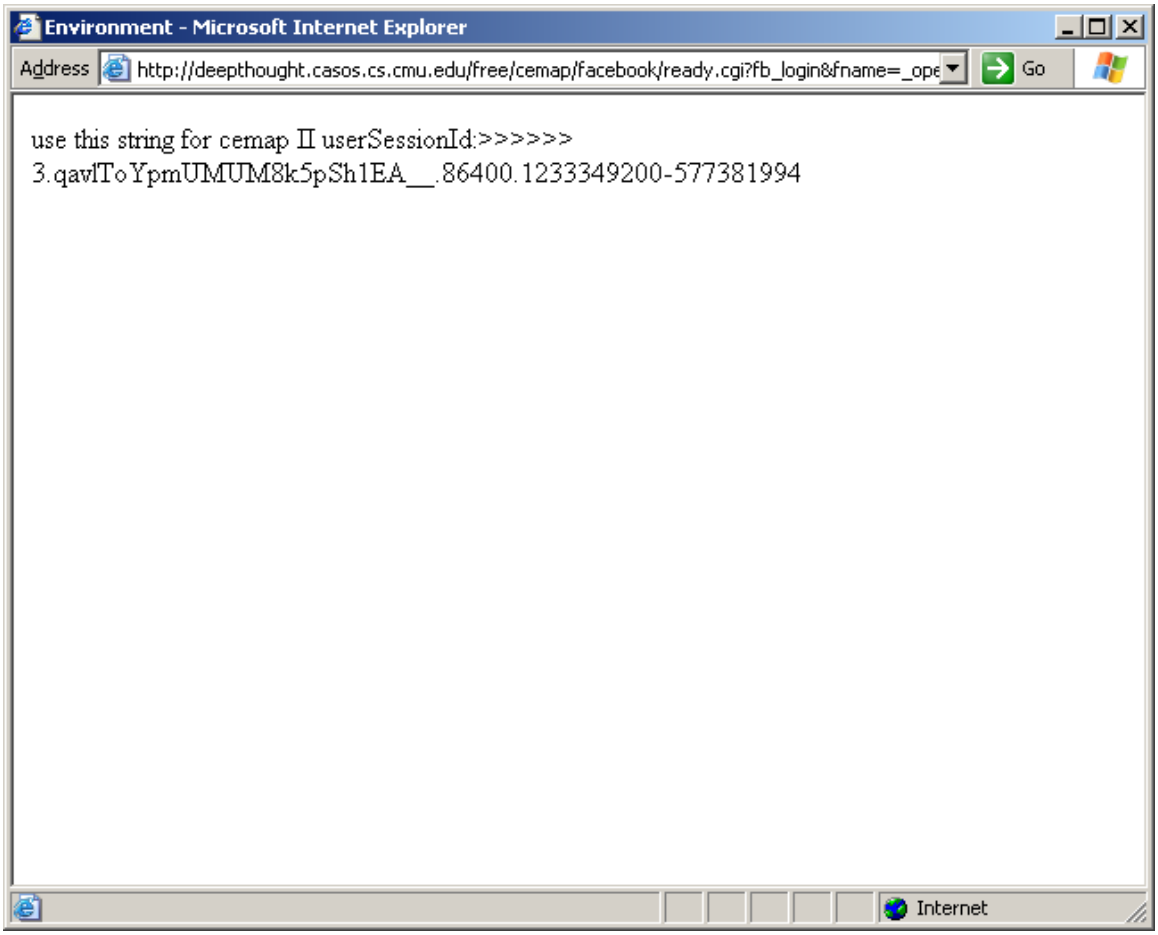
STEP 3: (If not already logged on) Log into the Facebook site using your Facebook email id and Facebook password.

If all is okay, you will then be presented with a simple page in the browser that looks like the picture below.

STEP 4: You want to make note of the lengthy string; in this case the string is

`3.qavlToYpmUMUM8k5pSh1EA__.86400.1233349200-577381994`

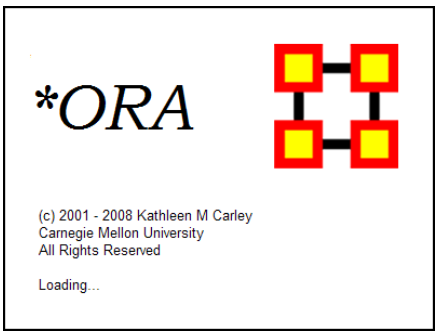
This long string of characters is a world-unique key that you provide to CEMAP and is then used by CEMAP, and only CEMAP, to access your Facebook temporality, and only while you are actually logged into Facebook.com. Keep this screen available to you when you are in the CEMAP process below. You will want to cut and paste this string rather than having to write it down, et cetera. If you lose this string, you can simply reperform this complete process, starting with Step 1.

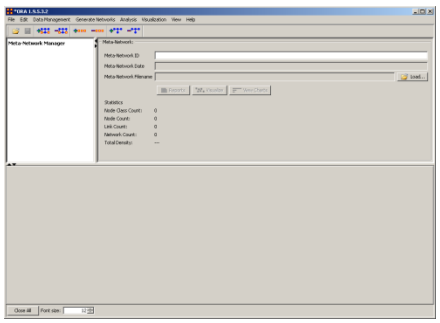


2.2 Performing the CEMAP Process

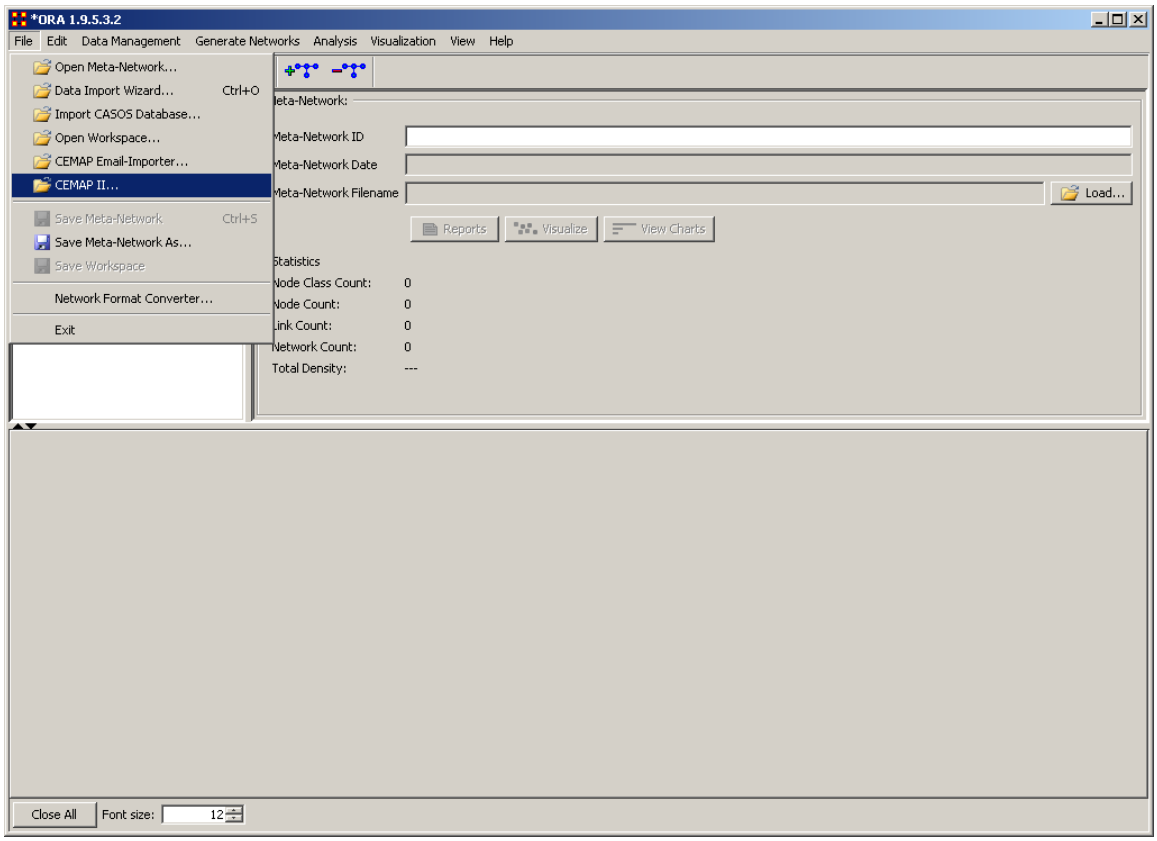
There are three essential steps the user ultimately must perform: (a) identify the input tablesets, (b) identify the output templates, and (c) set up mapping between the columns of the tables in the tablesets and the requirements of the various output template fields. However, these steps can be combined and hidden from a user such that a simplistic two-click process may be situated.

STEP 1: Start the ORA program via the Windows Program menu (other operating systems dependent process you use to start programs)

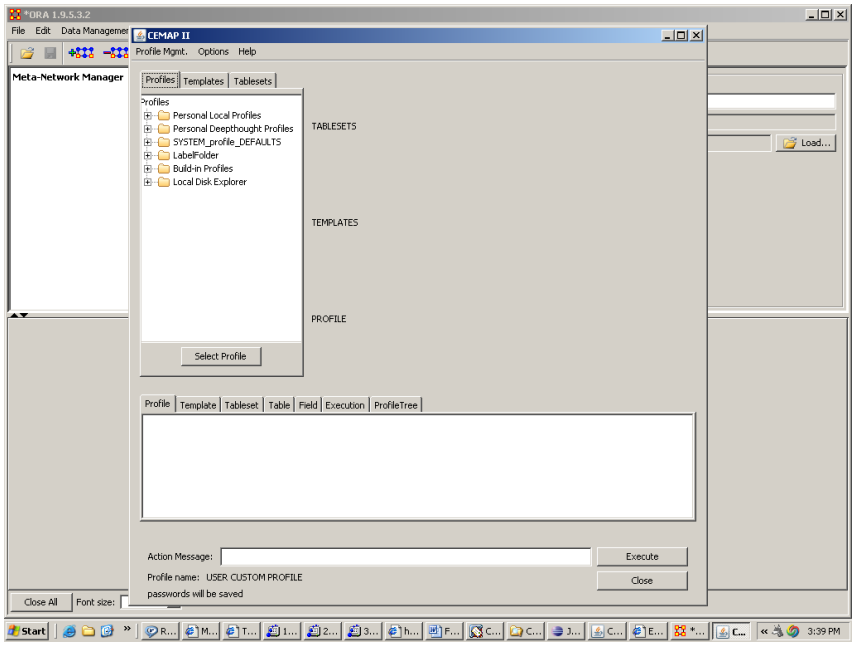




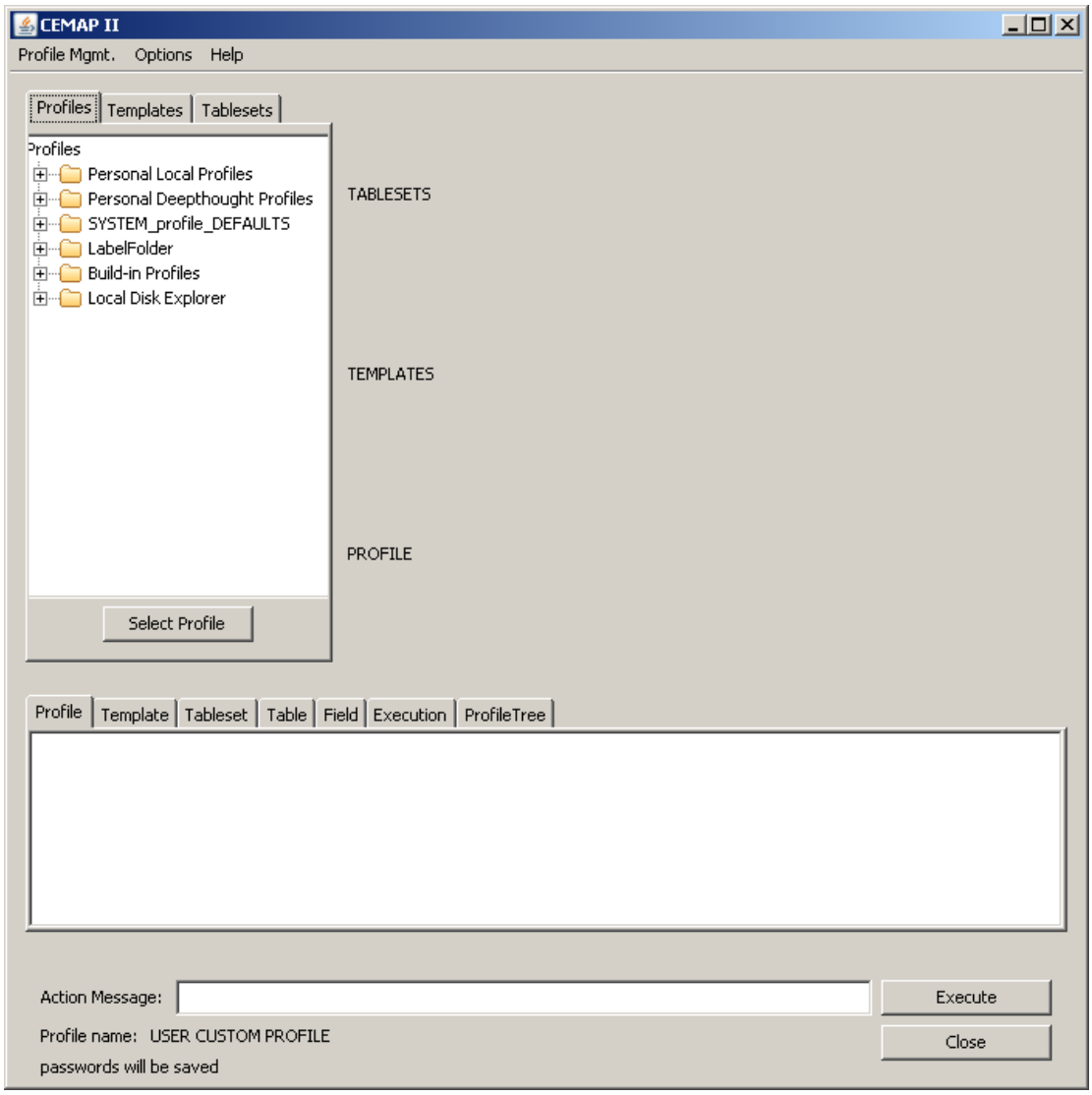
STEP 2: Select the File->CEMAP II menu option.



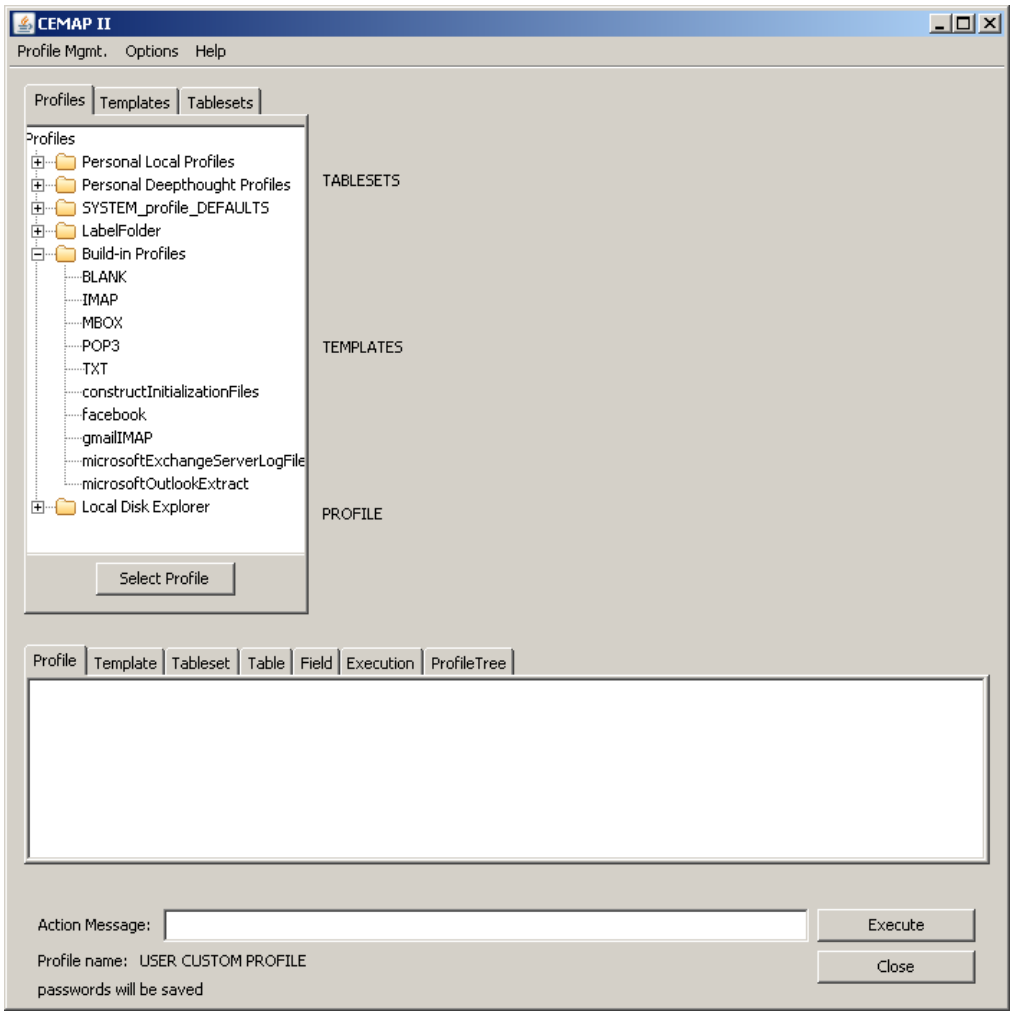
The CEMAP window pain will appear over top the ORA window, as shown below.



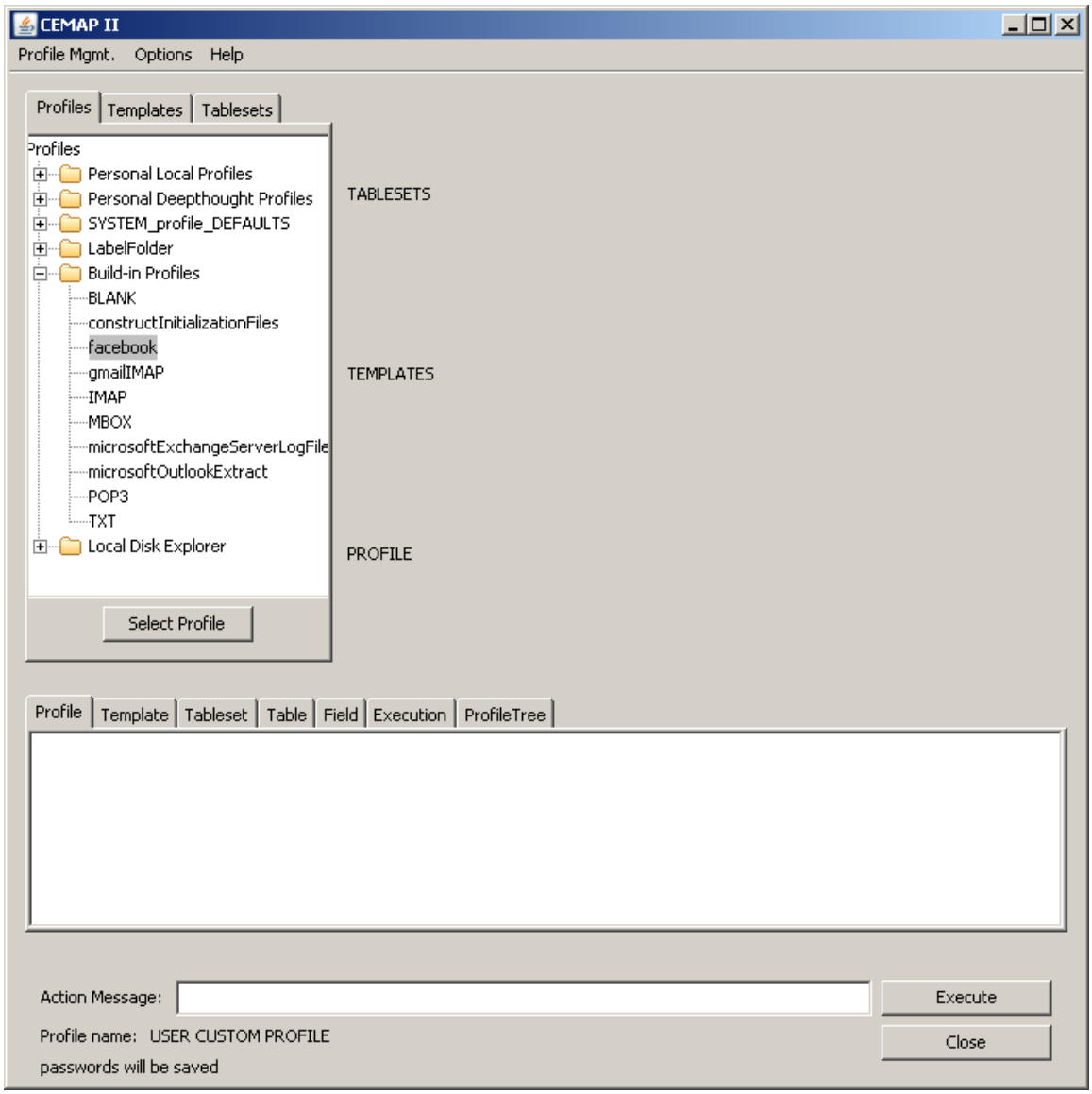
STEP 3: Left-click the "Build-in Profiles" item (Note: your particular profiles window may have different items shown. You will at least see "Build-in Profiles" and "Local Disk Explorer".)



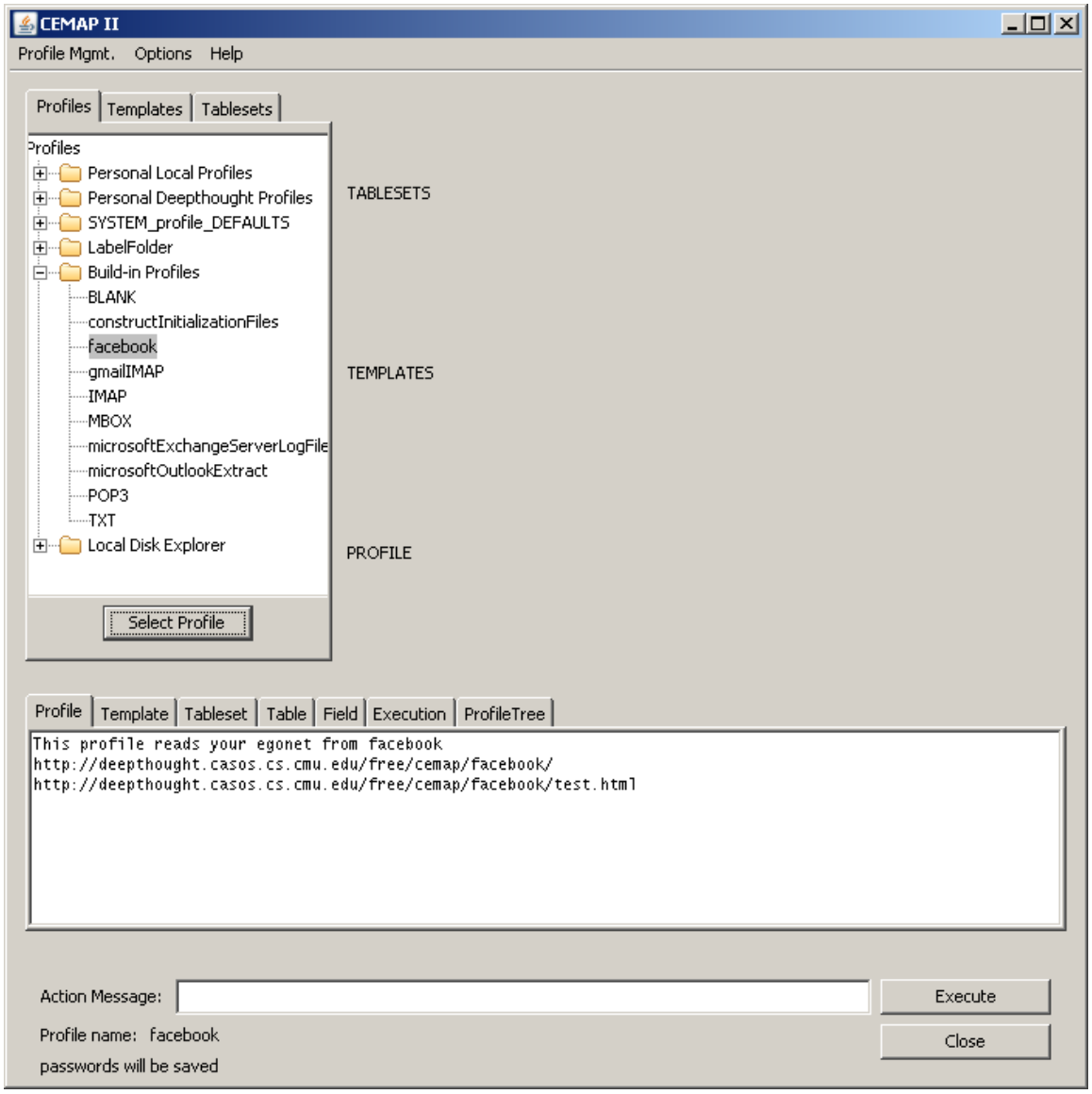
The item list will expand as shown below.



STEP 4: Left-click the “facebook” item so that it becomes highlighted.

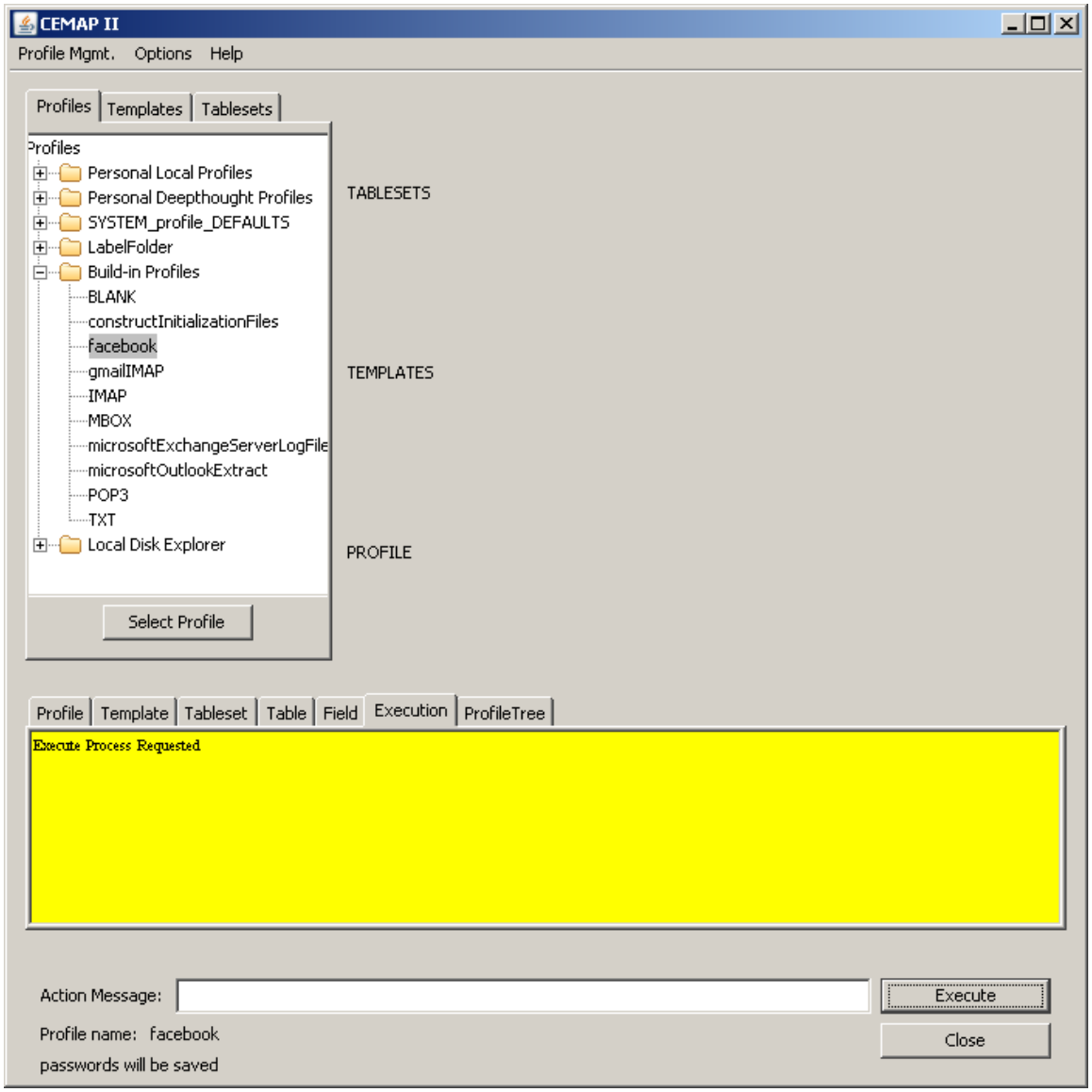


The Profile note panel will change and display some user notes and useful information.



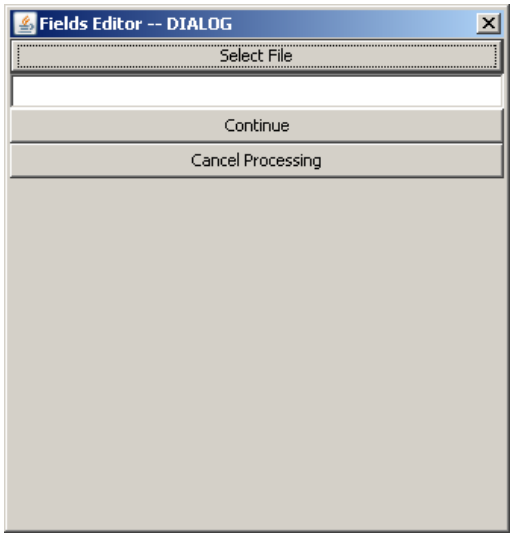
You have now loaded the special CEMAP Facebook profile into the CEMAP memory and you are ready to execute the process for collecting your data from Facebook for ORA.

STEP 5: Left click the "Execute" button. The Profile notes panel will turn yellow (to indicate that the execution has begun) and state that CEMAP is executing.

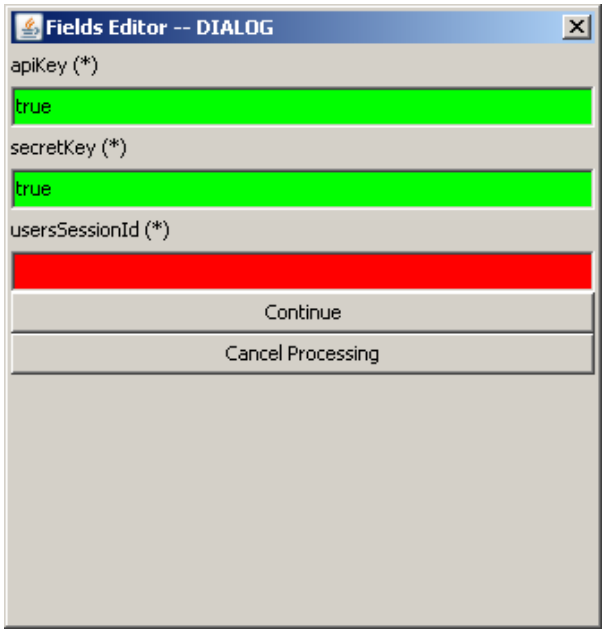


A small window will appear that is expecting a network output file name. This filename is not necessary because CEMAP is running in ORA. The output network will be loaded directly into ORA's memory, so you need not indicate a disk file location for the output. Optionally, you could Select File and indicate an output file name; the network file will still be loaded into ORA under either scenario. Complete the Select File process before taking the next step.

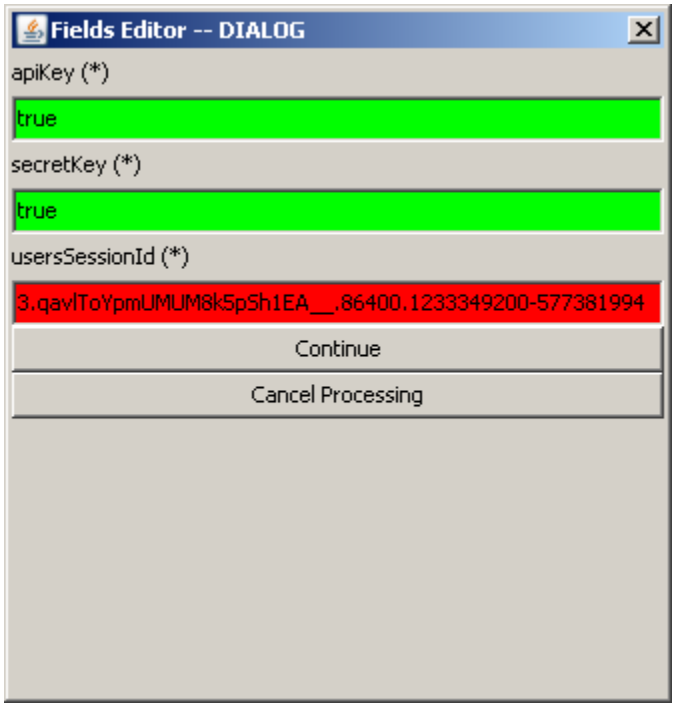
STEP 6: Left-click the Continue button.



The window will close and another small window will then appear, as shown below

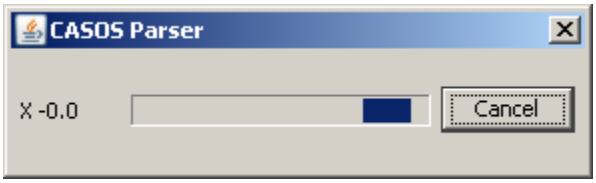


STEP 7: Enter the special access key string obtained in the Facebook authentication process (see prior sub-section of this document) into the third field in the window (the red field). Do not change the other fields. You can cut and paste this string from the browser window that provided the key earlier. Note: In some systems, you may need to use cntrl-v to actually paste the string into this field rather than righ-clicking the mouse.

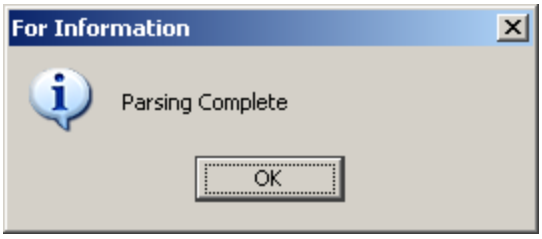


STEP 8: Left-click the "Continue" button.

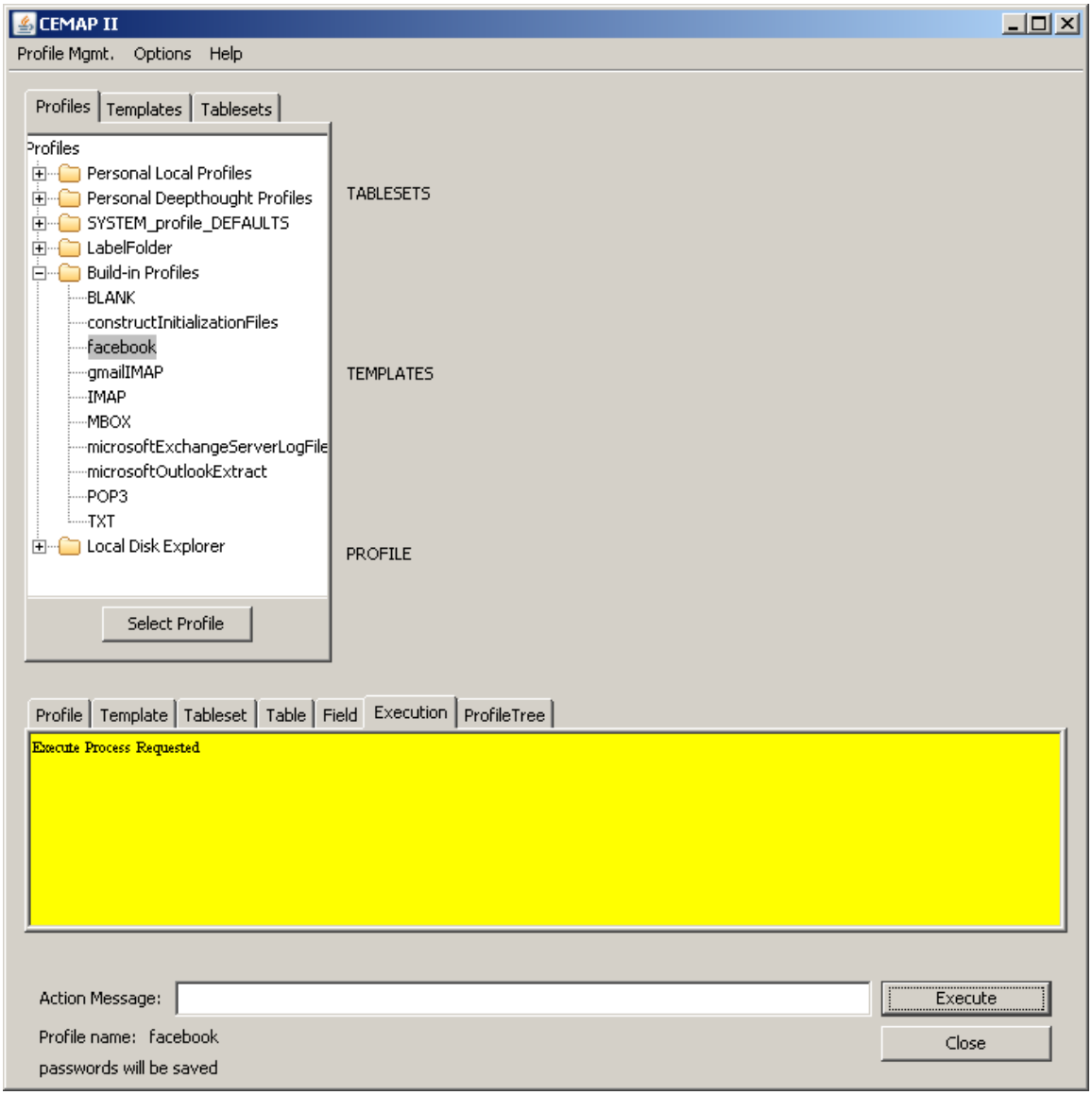
Now that you have supplied the information necessary for obtaining data from Facebook, CEMAP is now going to Facebook and asking for the information. A small window will appear that presented this process is underway. This window is shown below.



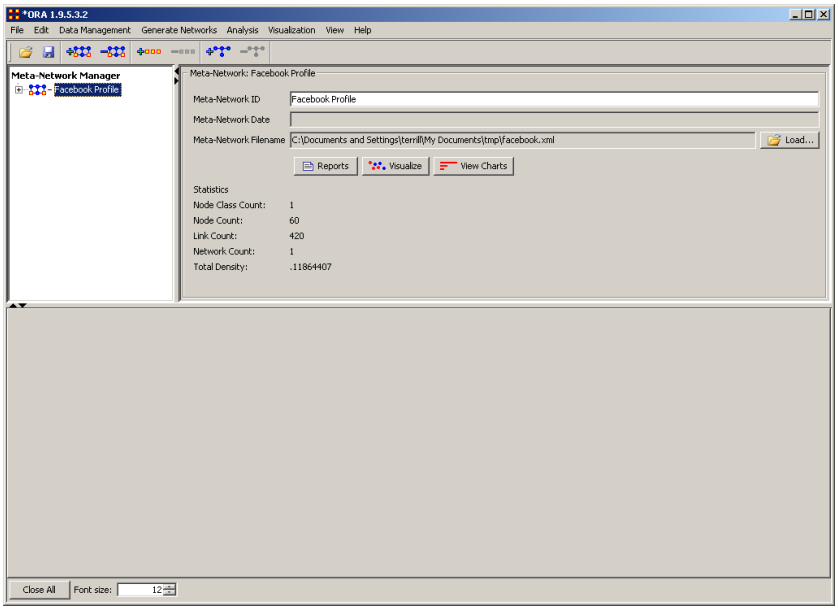
When the data collection process is finished. The progress window will close and an alert will be put on the screen that looks like below. If this process takes more than 1-2 minutes, there is a problem with the process and the user should try the process again. If still no success, technical support will need to get involved to diagnose the problem.



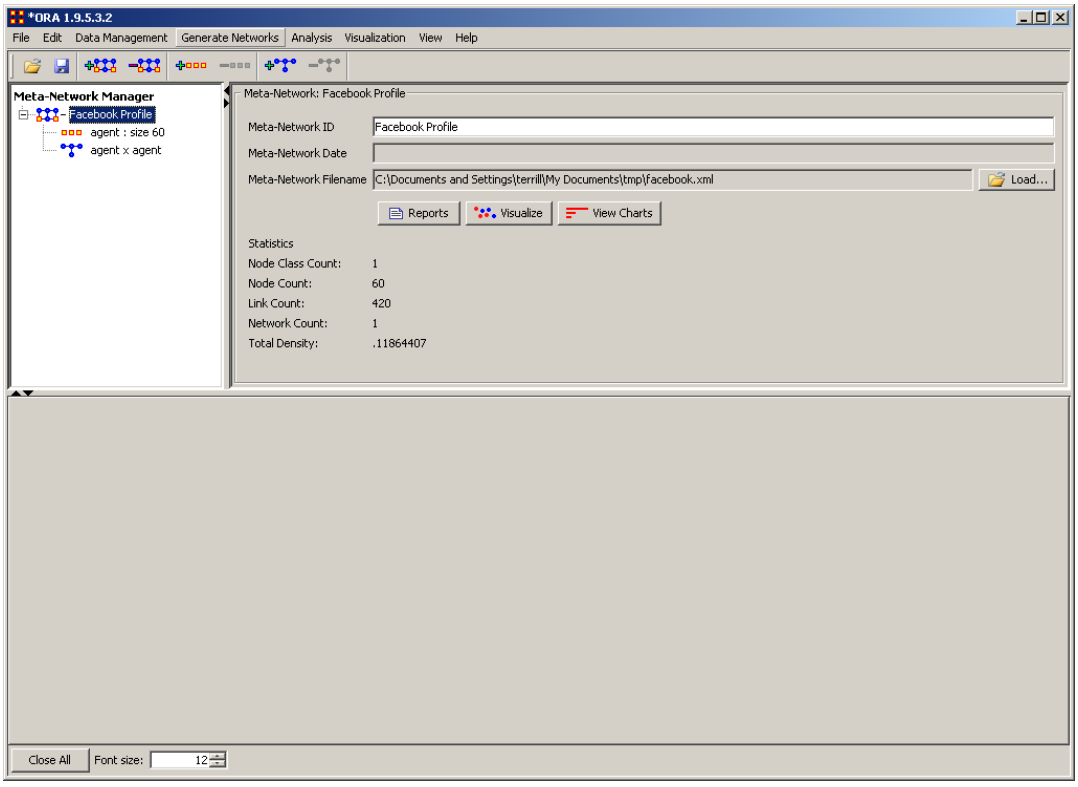
STEP 9: Left-click the "OK" button. The CEMAP window will remain.



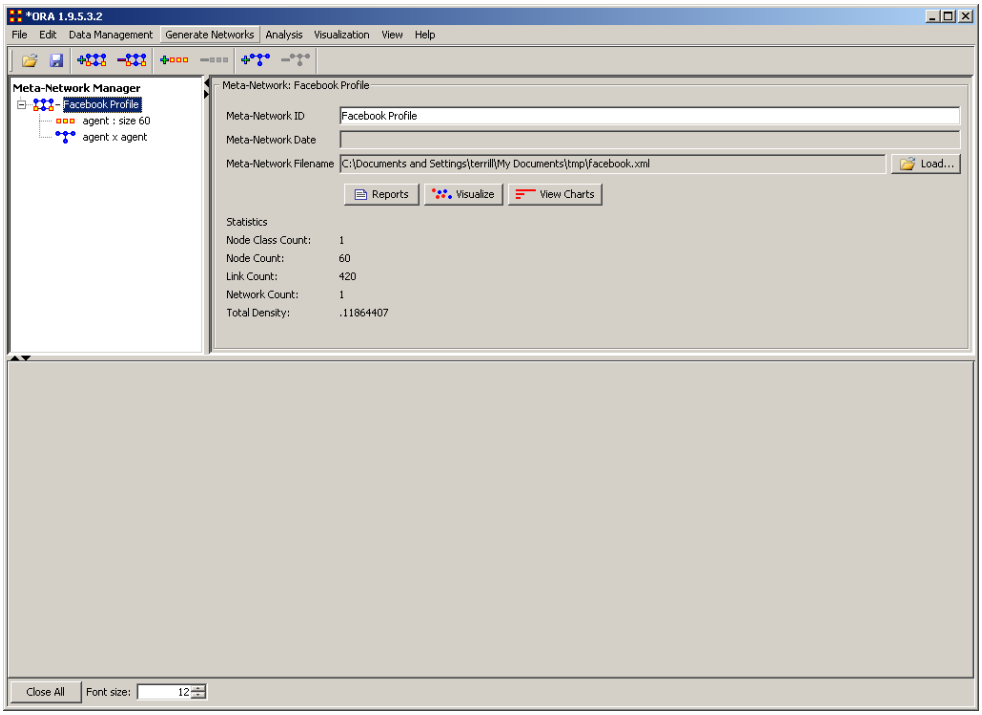
STEP 10: If all was successful, you will see the network data appear on the ORA data panel (which is behind the CEMAP window). Now, close the CEMAP window by left-clicking the "Close" button. You will now see the main ORA window as shown below.



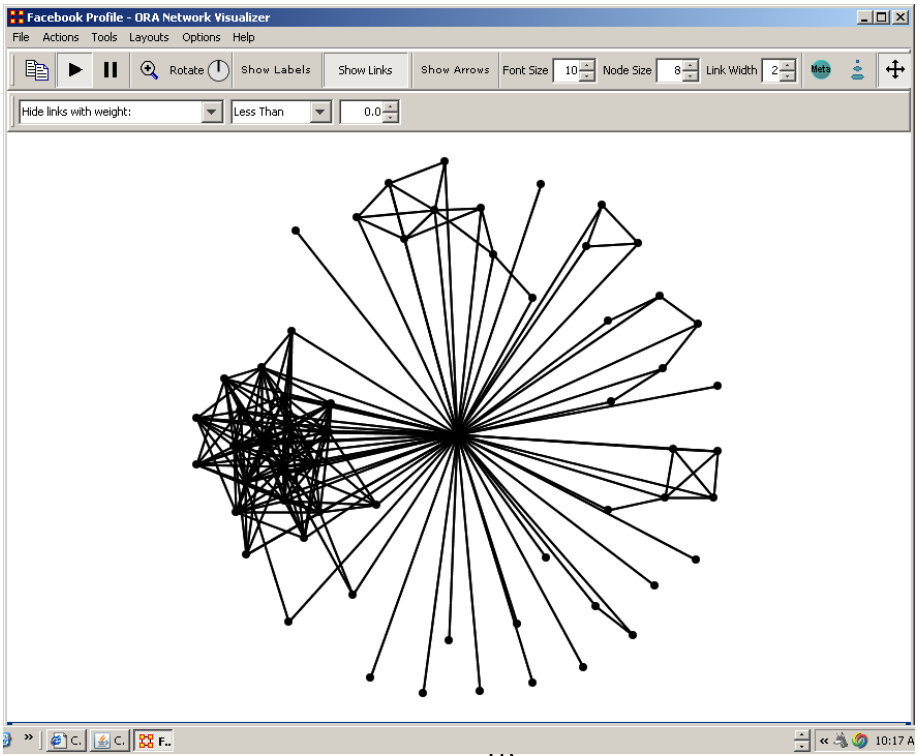
STEP 11: To see the structure of the new network file, left-click the tab for the "Facebook Profile" item and the nodeclass and network tabs will appear as can be seen below. This network data is fully loaded into ORA and you can now do what you like with it. If not familiar with ORA usage, try the online help for a quick tutorial and feature references.



STEP 12: Assuming you want to have a quick look at your Facebook network, left-click the "Visualize" button.



And, if all has gone well, congratulations, your Facebook ego-network should appear as is illustrated in the sample below. (For this picture, the node ids, or my friends names, have been removed; your names will likely appear automatically—depending on your current ORA settings.



3 Underlying Technology

The CEMAP Facebook ego-network data harvesting capability is possible from a interaction between two core technologies: CEMAP and the Facebook Developer Platform. The role of the Facebook Developer Platform is to make a subset of Facebook data available to third-party software, though the specific data that is available is merely a small portion of the data maintained by a user, and is strictly restricted by extending to only the data a single Facebook user can access. The role of the CEMAP software is to deal with the technical aspects of accessing this Facebook data and manipulating the data into a format necessary for use in ORA. This section provides some of the background of these two technologies as it pertains to the Facebook CEMAP profile.

3.1 CEMAP

The entire design of CEMAP II takes advantage of the notion that data stored in a tabular form has an inherent characteristic that each data item in a row is related in some comprehendible manner to any other data item in that same row. Since network data is essentially all about relational data, therefore, any pair of data items in the same row can form a link in a network. In short, given a table of data, each column can be the source for a DyNetML nodeclass, and each pair of data items in a row (or column by column) can be the source for a DyNetML network. It is this inherent characteristic of a data table that CEMAP takes advantage of. CEMAP provides features to convert real-world data into a tabular form and to describe DyNetML and AutoMap formats in a simplistic manner. CEMAP then facilitates the mapping of the tabular data into the output format according to the desires of the analyst. Essentially, CEMAP conveniently separates the technical task of formatting real-world data into a generic, consistent and easy to understand table from the network-centric vantage point that an analyst holds, then provides a simple mapping mechanism to join the two representations.

There are four primary component parts to the CEMAP architecture and user-design that even than the most basic user (a user that only is concerned with completed profiles) of CEMAP should be somewhat familiar with. A *template* is a description of an output format from CEMAP. These outputs most often consist of DyNetML templates, but there are numerous output templates that describe other output formats. Most of the output templates can be customized in some manner, according to the designer of the template. For example a DyNetML template allows the user to add, delete or change the characteristics, of a nodeclass, or network, among other several features of ample flexibility. The user has a great amount of flexibility in the flat file template as they are not constrained by the strict requirements of DyNetML. For example, a template can be customized to allow for the user to create a CSV file to import the output data into an Excel spreadsheet. This is merely by-product of the CEMAP II feature set, which is designed expressly for the CASOS software suite.

A *tableset* is a collection of tables that logically belong together either in the manner in which the original data is stored, or from the perspective of the user. For example, a collection of SQL statements involving a single database host and password,

etc, can situation for a tableset, with each table corresponding to an SQL statement, but with the same characteristics as the other tables within the same tableset. A tableset translates the physical data source into a row-column oriented table that can be mapped with one or more other tables or template. A tableset takes much., most, or all of the complexity out of a user reaching an original data source for their subsequent ORA/AutoMap analysis.

A *profile* is a file whose main role is to keep the completed mappings of a tableset and template(s). There are usually the three logical aspects of a process fully defined within it. This is a CEMAP II file that has a set of template(s), a set of tableset(s), and the mapping between the two. A mapping without a tableset is impossible. A mapping is always associated with a template. The profile can have any, all, or none of the fields within the template, tableset completed, or have the matching completed. A profile might correspond with a routine task that the user has to perform often or periodically. The profile is meant to "remember" all of the details pertinent to the CEMAP II task, also some details can be purposely left unsatisfied by the creator of the profile (the actual name-location of the output file, perhaps, among other things like high-confidential passwords and such).

A resource is a mapping of a physical file, data or CEMAP specific file (profile, tableset, template, or other resource file), that the user has access to. This resource can reside anywhere reachable to the user's computer, e.g. local disks, the Internet, etc. The function of a resource is to simplify the reaching of the file to the user. That is it saves the user from remembering long URL's or file paths, etc.

3.2 Facebook's Developer Platform

The Facebook Developers Platform provides APIs that third-party software programs can use to access user data. For complete information on the platform, visit Facebook.com for the latest and detailed information. This section speaks to only the APIs that CEMAP uses in the Facebook CEMAP profile.

CEMAP uses the REST-like API interface to Facebook. HTTP POST requests are constructed and sent to Facebook, which in turn returns the results in an XML package that gets parsed into the table-row format necessary for CEMAP processing. CEMAP first sends a `getUID` request to obtain the `userId` for the user. It hen makes a `getInfo` request for each friend to obtain their name from each friend id. Next a `friends_get` request is made for that user to obtain the list of friends for the user; this results in the ego-network data. Then `the friends_areFriends` request is made with a full list of pairs of the friends with results in identifying the ties between pairs of the friends. This is as far as Facebook will permission the APIs to reach in the egonetwork, so the CEMAP process stops there.

The process to obtain the user's session Key string is a webpage that interfaces with Facebook using Javascript. This page executes the Facebook `init` and `getSessionState` methods. Which, if and when successful, returns a response via an HTTP GET call to a CGI script. When this CGI script is executed on the server, it parses the URL string that made up the call to itself looking for the `session_key` keyword and its paired value from the GET string. The string is then displayed in the browser window for the user.

4 Future Work

The CEMAP Facebook profile is severely limited in what data it can access from Facebook and this is entirely at the discretion of Facebook. CEMAP can only provide what data is permitted by the Facebook host, via APIs. There are a few fields for individuals that are currently available from Facebook that is not yet part of the CEMAP implementation. For example, users' hometown-location, gender, and their location (city and state) are available; this attribute data will be made part of the CEMAP implementation in the future, assuming that Facebook keeps these data available to developers via the APIs.

5 References

Carley, Kathleen. (1991). A Theory of Group Stability. American sociological Review, 56, 331-354.

Carley, Kathleen, Diesner, Jana, Reminga, Jeffrey, Tsvetovat, Max. (2004). Interoperability of Dynamic Network Analysis Software.

Carley, Kathleen & Reminga, Jeffrey. (2004). ORA: Organization Risk Analyzer. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-106.

Davis, George B., Olson, Jamie, & Carley, Kathleen M. (2008). OraGIS and Loom: Spatial and Temporal Extensions to the ORA Analysis Platform. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISR-08-121.

Diesner, Jana & Carley, Kathleen. (2004). AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-100.

Facebook (2009). www.facebook.com Retrieved on 29 January, 2009.

Frantz, Terrill & Carley, Kathleen. (2008). *CEMAP II: An Architecture and Specifications to Facilitate the Importing of Real-World Data into the CASOS Software Suite.* Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-08-130.

iStrategyLabs, www.istrategylabs.com (2009). *2009 Facebook Demographics and Statistics Report: 276% Growth in 35-54 Year Old Users*. Retrieved on 29 January 2009, http://www.istrategylabs.com/2009-facebook-demographics-and-statistics-report-276-growth-in-35-54-year-old-users/.

Wikipedia (2009). Facebook. Retrieved on 29 January, 2009, http://en.wikipedia.org/wiki/Facebook.

Tsvetovat, Max & Reminga, Jeffrey & Carley, Kathleen. (2003). DyNetML: Interchange Format for Rich Social Network Data. NAACSOS Conference 2003, Day 2, Electronic Publication, Pittsburgh, PA.