SC/Gen and the Social Networking Report

User Manual, SC/Gen v1.0.14
and Social Networking Report v1.0.5

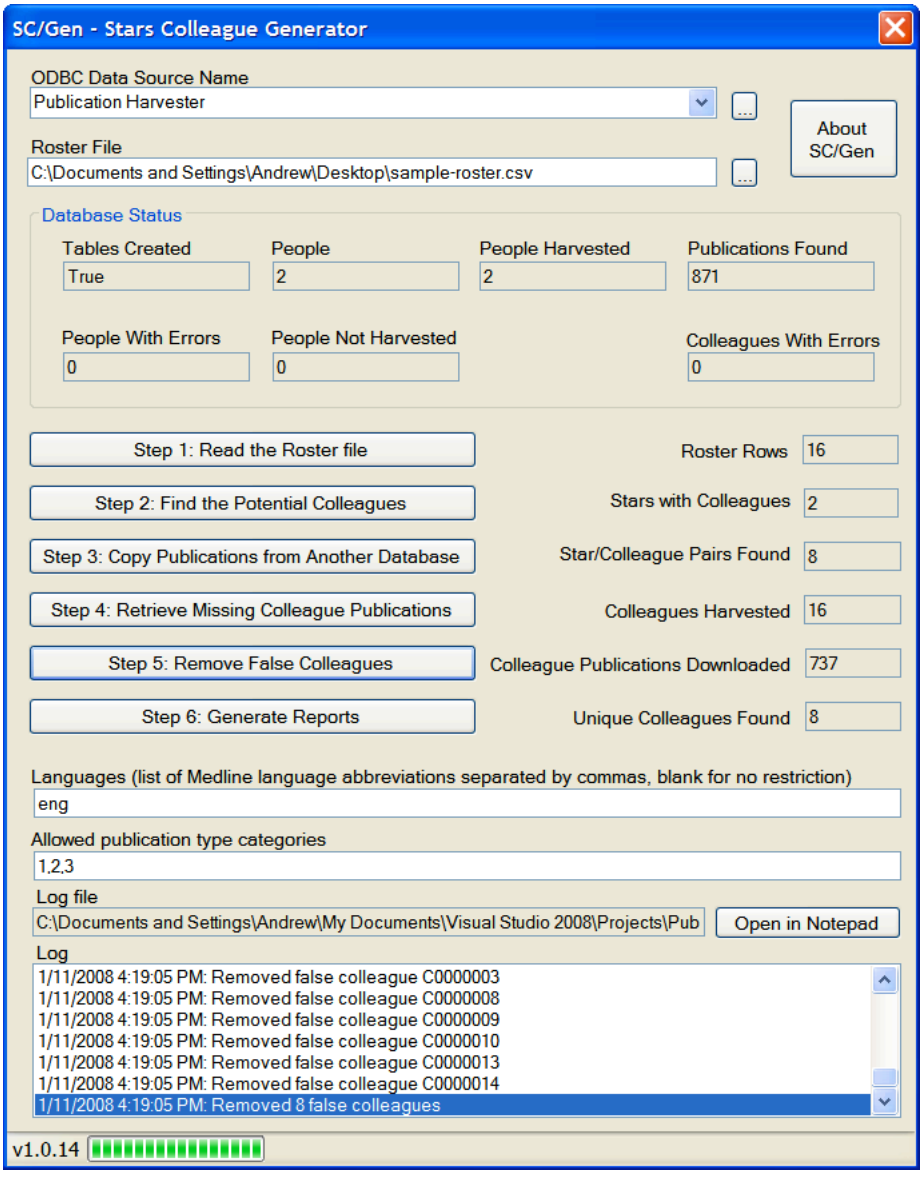


Table of Contents

1 Introduction

1.1 Purpose

The purpose of this document is to serve as a guide to people who want to use the SC/Gen software and the Social Networking Report software. It should give them all of the information necessary to install, configure and use the software.

1.2 Scope

This document contains step-by-step instructions to show users how to install, configure and use the SC/Gen software and Social Networking Report software on a machine running Windows XP. It covers:

- Installing the SC/Gen software
- Preparing the input files
- Using SC/Gen to generate colleagues from a Publication Harvester database
- Using SC/Gen to generate reports
- Using the Social Networking Report software to generate a social network report from two databases generated by SC/Gen

1.3 System Overview

The purpose of the SC/Gen is to search through a database created by the Publication Harvester project to identify potential colleagues of each star, where a colleague is a person who coauthored a particular publication with a star. It is then used to gather publications for each colleague, and generate reports for analysis of the colleagues in the database.

The purpose of the Social Networking Report is to search through two databases that were created by SC/Gen for the same set of stars. The first database (called the regular database) is used to connect colleagues to their stars to form a first-degree social network. The second database (called the square database) is used to connect those stars to other stars they coauthored with – this is the second-degree social network.

1.4 References

The SC/Gen software requires a database that was created using the Publication Harvester software. This manual does not explain the use of that software – the manual and specification of the Publication Harvester software can be found at http://www.stellman-greene.com/PublicationHarvester/

This manual does not go into detail about what constitutes a colleague, the formats of the input files and the reports, or the structure of the database. All of that information can be found in the software requirements specification (SRS): http://stellman-greene.com/SCGen/SRS_Colleagues.doc

2 Installation

This section describes how to install the SC/Gen software.

2.1 Before you start…

SC/Gen is built to operate on a database that was created by the Publication Harvester. The SC/Gen software also requires .NET Framework 3.0, MySQL 5.0 and MySQL/ODBC 3.51. You can find installation instructions for this in the manual for the Publication Harvester, which can be downloaded from http://stellman-greene.com/PublicationHarvester/Publication%20Harvester%20Manual.doc

2.2 Install the SC/Gen and Social Networking software

Download the latest version of the SC/Gen and Social Networking installers from http://stellman-greene.com/SCGen/bin/latest/, unzip it and run setup.exe to install the software. Once each installer is finished, it will run the software it installed. It will appear in the Start menu listed under "Publication Harvester". In addition, there are sample files that can be downloaded from the SC/Gen website:

- sample-roster.csv -- sample input roster file that contains potential colleagues
- sample-JIFs.xls -- sample JIF file for generating reports
- sample-square-roster.csv -- sample square roster file generated to match the Publication Harvester sample input file
- sample-input.xls -- sample input People file you can use with the Publication Harvester
- sample-colleagues.txt -- sample colleague file used by the Social Networking Report

2.3 Overview of the SC/Gen software

The purpose of SC/Gen is to identify the colleagues of people who were "harvested" using the Publication Harvester. The Publication Harvester starts with a list of people, finds each person's publication citations using PubMed, and saves them in a MySQL database. SC/Gen picks up where Publication Harvester left off by reading the data for the people and their publications from the database:

1. The software reads a roster file that contains information for potential colleagues.
2. The software generates colleagues by searching through the publications for each person in the Publication Harvester database and comparing the coauthors to the people in the roster file. A person's colleague is someone who coauthored a publication with that person and appears in the roster.
3. The publication citations for each colleague are added to the database, either by copying them from another "pre-harvested" database or by retrieving them from PubMed.
4. There will be some false or spurious colleagues that were found in step 2 who, upon finding their publications, don't list the original person as a coauthor. Those false colleagues are removed from their colleague lists.
5. Reports can be generated for statistical analysis.

3 Using the SC/Gen software

Now that SC/Gen has been installed, it can be used to generate colleagues. Before you can do that, you'll need to use the Publication Harvester to create a database. Once you have that database, SC/Gen will search through it, generate colleagues, and create the reports.

3.1 Specify the Roster File

The SC/Gen software uses a roster file to select potential colleagues. To do that, it uses a roster file. It's a CSV file that can contain a roster of scientists, one row per person. It can also contain a smaller or different set of people. The CSV file contains the following columns:

- setnb (text [length=8]): identifier for the person
- fname (text [length=20]): first name
- mname (text [length=20]): middle name
- lname (text [length=20]): last name
- match_name1 (text [length=20]): PubMed-formatted name
- match_name2 (text [length=20]): PubMed-formatted name (optional)
- search_name1 (text [length=20]): PubMed-formatted name
- search_name2 (text [length=20]): PubMed-formatted name (optional)
- search_name3 (text [length=20]): PubMed-formatted name (optional)
- search_name4 (text [length=20]): PubMed-formatted name (optional)
- query (text [length=244]): A search query which will be used to retrieve publications from Pubmed

The matchname1 and matchname2 columns are used to match the person in the roster to a person's publication. If either of these names shows up in the list of authors in a person's publication, then the person is a colleague. (matchname2 can be empty, in which case the software only looks for matches against matchname1.)

The searchname1 through searchname4 columns are used to look in the results of a PubMed query to find a colleague's publications. If any of those names matches a name in the author list of a returned citation, then that colleague is an author of the publication. (searchname2 through searchname4 can be empty; the software will only search on the provided names.)

Each of the name columns contain a name in the same format as the author list in a PubMed citation (e.g. for Robert E. Elston, name1 might contain "ELSTON RE").

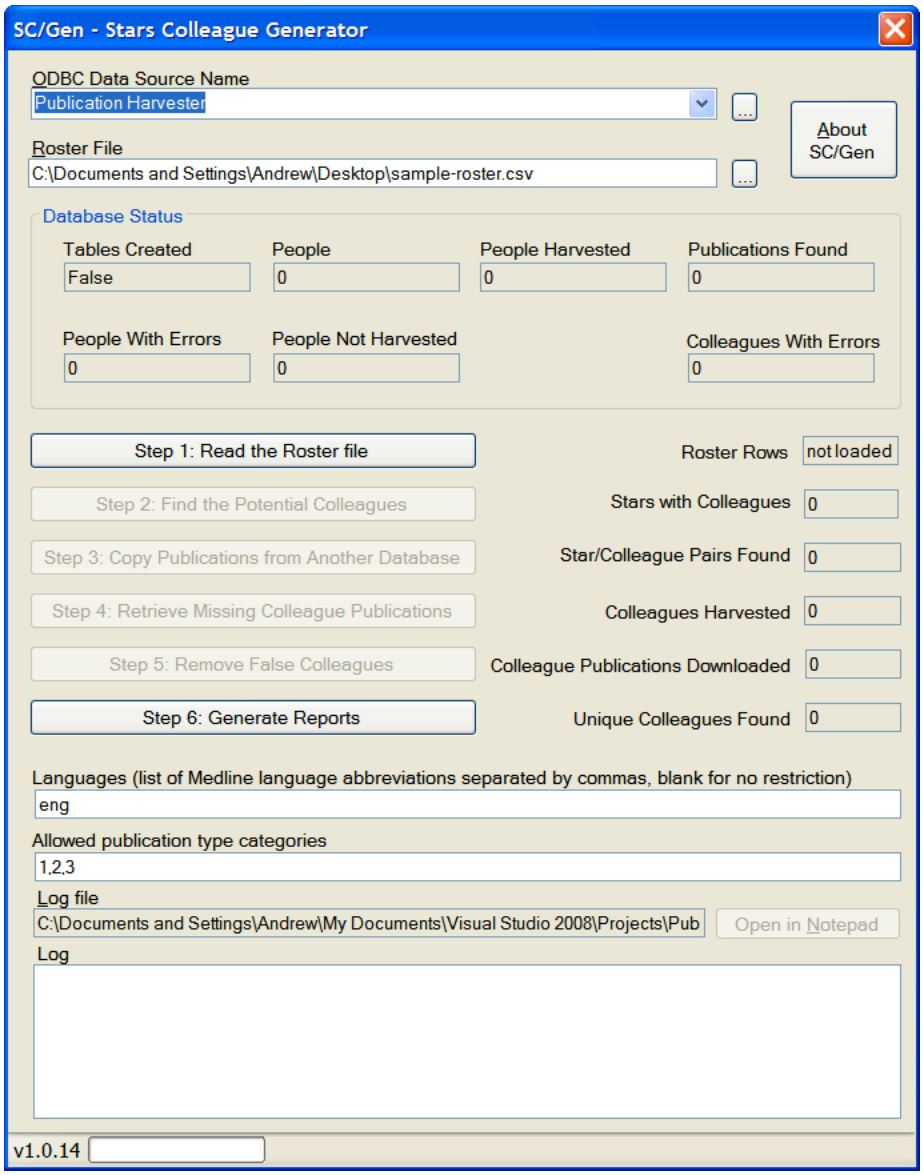
The search_name1 through search_name4 columns are used to search PubMed and retrieve citations (in the same way as in the People file – see the Publication Harvester SRS).

A sample roster file can be downloaded from the following URL:
http://stellman-greene.com/SCGen/sample-roster.csv

3.2 Generate Colleagues

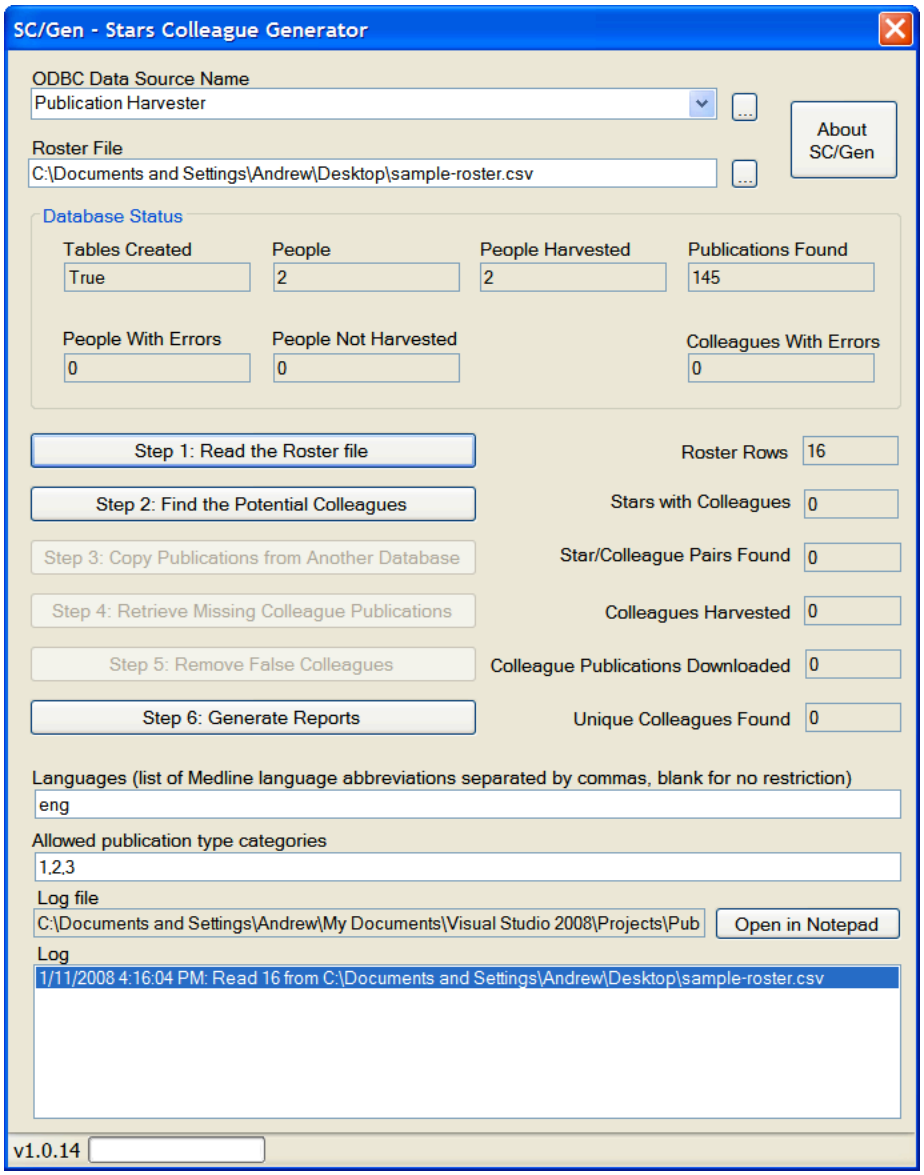
Before you can generate colleagues, you'll need to create and populate a database using the Publication Harvester. Once that database is created and populated, start the SC/Gen software and select the ODBC data source name you used with the Publication Harvester. (If you click the "…" button next to that field, it will pop up the ODBC Data Source Administrator.)

Click on the "…" button next to the "Roster File" field and browse to the Roster. Your SC/Gen window will look like this:



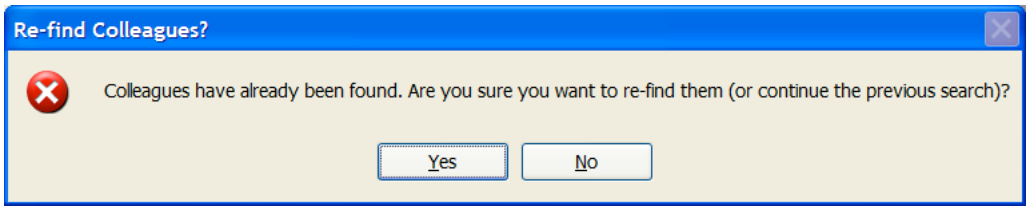
Click the button labeled "Step 1: Read the Roster file". SC/Gen will read the roster file into memory. It also creates an XML file in the same folder that contains the same information as the roster. For a very large roster file, it's faster to load the XML file than it is to load the CSV file. The XML file will have the same name as the CSV file, with ".xml" added to the end ("sample-roster.csv.xml"). The number of rows in the roster will be displayed in the "Roster Rows" box.

Once the roster is read, the button labeled "Step 2: Find the Potential Colleagues" will be enabled.

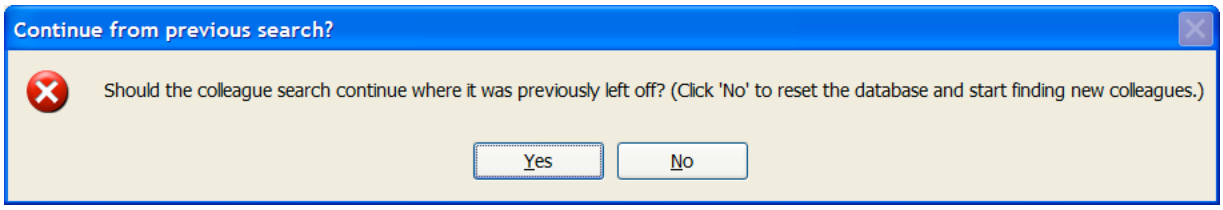


Click the "Step 2" button to tell SC/Gen to read the database and find the potential colleagues. The software will add additional tables to the database to hold the colleague information.

If you click the Step 2 button on a database that already contains colleagues that were found, it'll display a warning box:

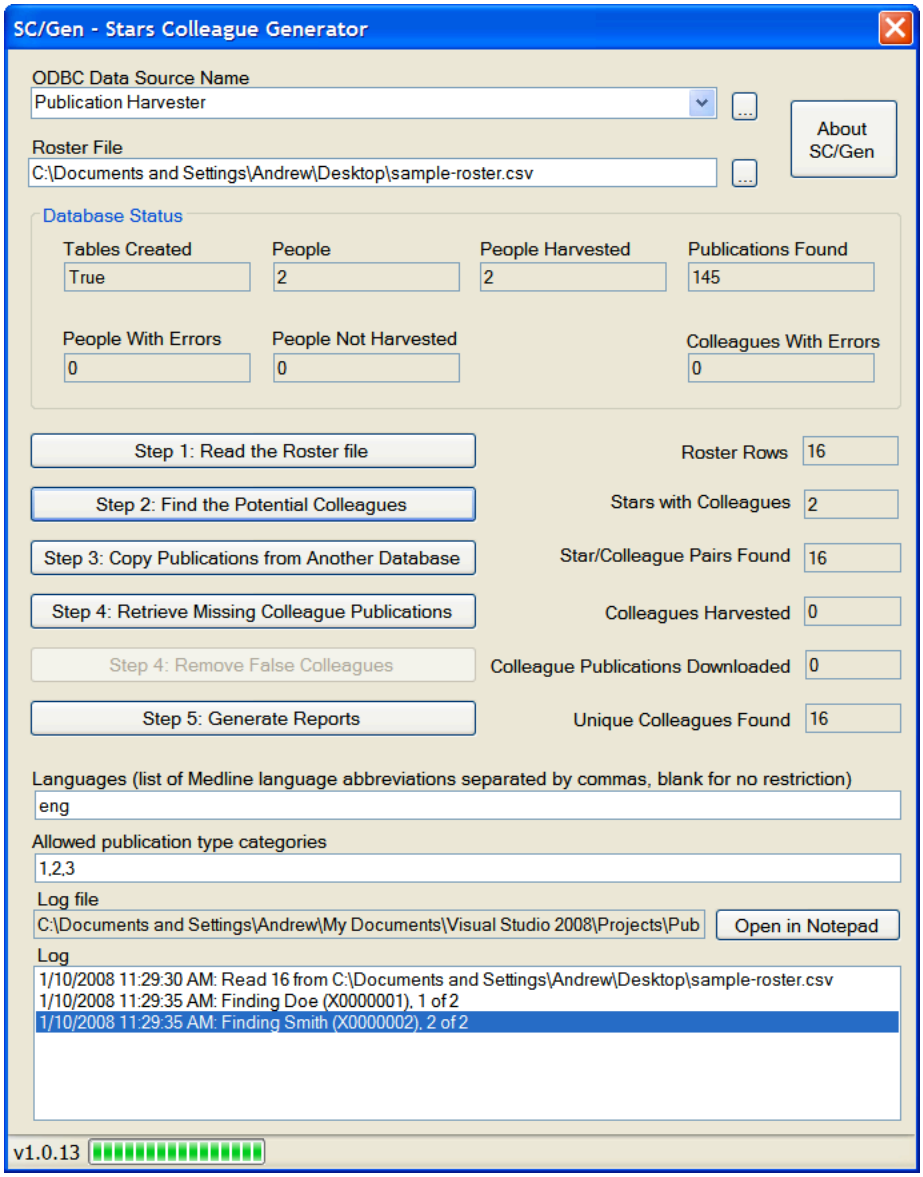


If you click "Yes", you can choose to either continue the previous colleague search or reset the database and find new colleagues:



Those two windows will only appear if there are already colleagues in the database. If you're using a clean Publication Harvester database that's never had colleagues generated, those windows will not pop up.

Once the potential colleagues are found, the system will show data in three additional boxes. "Stars with Colleagues" contains the number of distinct stars that it finds in the StarColleagues table – that's the number of stars that have at least one colleague. The "Star/Colleague Pairs Found" contains the number of pairs of stars and colleagues. And the "Unique Colleagues Found" box contains the total number of unique colleagues across all stars in the entire database.



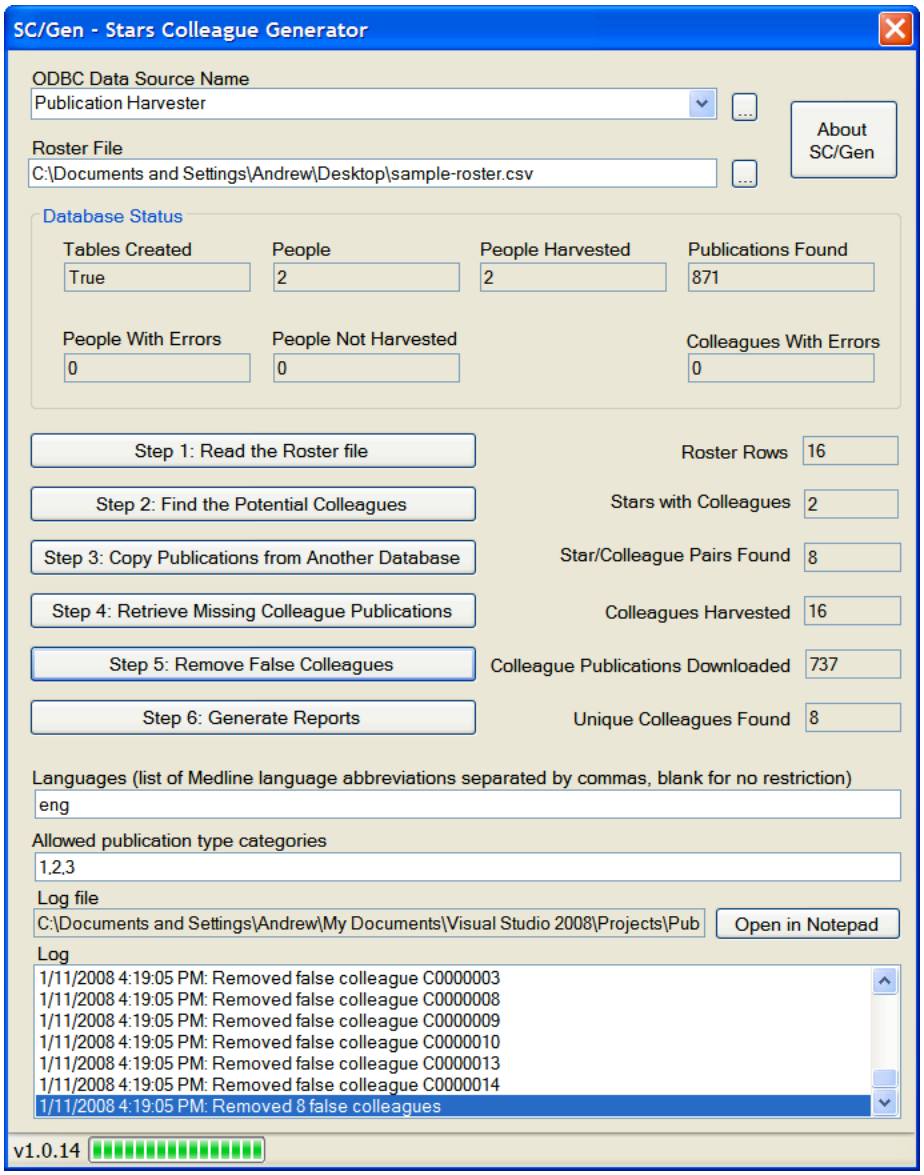
Once the potential colleagues are found and added to the database, each colleague's publications need to be harvested. There are two ways to do that. If you have a lot of colleagues whose publications you'll be repeatedly harvesting, you can create a separate Publication Harvester input file with those colleagues and add them to a separate database on the same MySQL server. Once you've got that, you can click the "Copy Publications from Another Database" button, which brings up this window:

Select the database that contains the publications to copy. When you click the "Copy Publications" button, the program will copy the publications from one database to the other. It will mark the colleagues for which the publications were copied as "harvested" – that way, when you retrieve the publications later for the remaining colleagues, the system won't take the time to download and process those colleagues' publications. It's much faster to copy publications from another database than it is to download them from PubMed, so this can save a lot of time. You can copy publications repeatedly from several other databases.

Once you've copied all of the publications, click "Step 3: Retrieve Colleague Publications". This does exactly the same thing for the colleagues as the Publication Harvester does for the stars – it downloads the publications from PubMed for each colleague.

Note: If this process is interrupted or the machine is shut down, you can restart the harvest later. The SC/Gen software will pick up where it left off, without losing any data.

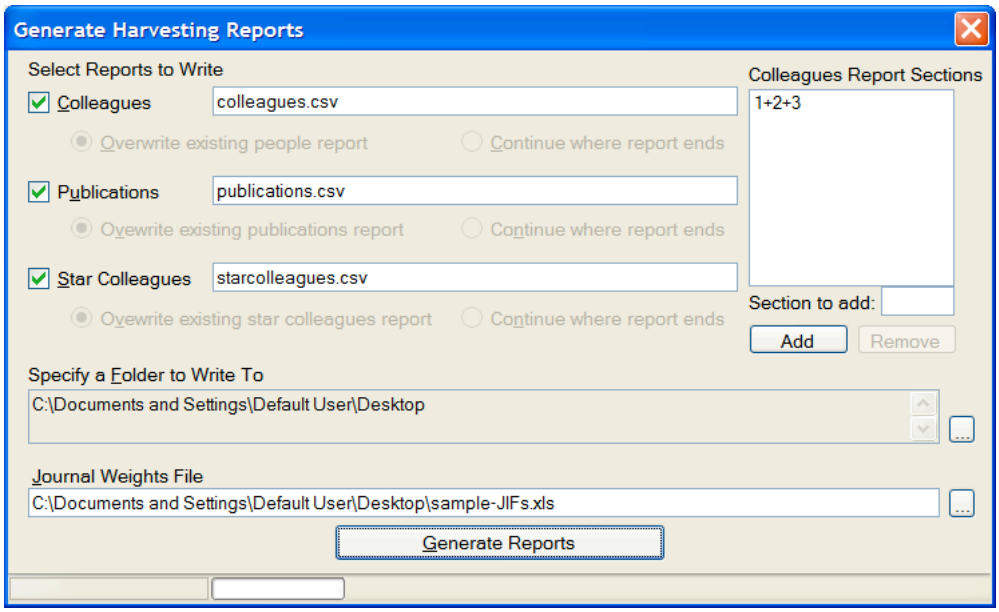
Once the colleagues are harvested, it updates the "Colleagues Harvested" box (which contains the number of unique colleagues that have had their publications downloaded) and the "Colleague Publications Downloaded" box (which contains the total number of unique publications – if two colleagues coauthored the same publication, it will only be counted once).



There are many cases where you'll find roster matches that don't actually represent real colleagues. For example, there could be two John Smiths with different PubMed queries. If the star has a publication with author "SMITH J", both roster rows will be matched. But once the colleagues' publications are harvested, it's possible to check each colleague's publication list against the star's publication list. If there are no publications in common, then the colleague is a "spurious colleague". Step 4 removes those false colleagues from the database and updates the “Stars with Colleagues” box and the “Star/Colleague Pairs” box. (The colleagues are not entirely removed from the database; they are just disassociated from the stars. That way, if one person is a colleague to two stars, his publications remain in the database.)

3.3 Generate the colleague reports

Once all colleagues are harvested and false colleagues are removed, click the Step 5 button to generate the reports. It brings up this dialog box:



You can specify the names of the three reports using the boxes. They're all generated in the same folder; use the first "…" button to specify the folder to write to. If any of the files exist in the folder already, use the "Overwrite existing report" or "Continue where report ends" radio buttons to specify whether to overwrite or continue the report.

You'll need to specify a Journal Weights (JIF) file for the Colleagues report. This is the same JIF file that was used with the Publication Harvester. Click the "…" button next to the "Journal Weights File" box to locate it.

The **Colleagues** report is exactly the same as the People report in the Publication Harvester. By default, it only contains summary rows for publication type "bins" 1 + 2 + 3:

Field	Type	Description
setnb **(key)**	Text	Colleague unique identifier
year **(key)**	Number	Year of publication
pubcount	Number	Total nb. of pubs in year, bins I+II+III
wghtd_pubcount	Number	Weighted total nb. of pubs in year, bins I+II+III
pubcount_pos1	Number	Total nb. of pubs in year, bins I+II+III, 1st author
wghtd_pubcount_pos1	Number	Weighted total nb. of pubs in year, bins I+II+III, 1st author
pubcount_posN	Number	Total nb. of pubs in year, bins I+II+III, last author
wghtd_pubcount_posN	Number	Weighted total nb. of pubs in year, bins I+II+III, last author
pubcount_posM	Number	Total nb. of pubs in year, bins I+II+III, middle author
wghtd_pubcount_posM	Number	Weighted total nb. of pubs in year, bins I+II+III, middle author
pubcount_posNTL	Number	Total nb. of pubs in year, bins I+II+III, next-to-last author
wghtd_pubcount_posNTL	Number	Weighted total nb. of pubs in year, bins I+II+III, next-to-last author
pubcount_pos2	Number	Total nb. of pubs in year, bins I+II+III, 2nd author
wghtd_pubcount_pos2	Number	Weighted total nb. of pubs in year, bins I+II+III, 2nd author

You can add additional sections for additional publication types by typing the section to add and clicking the "Add" button. For example, you can specify that the report also contain information that only includes data from bin #2 by adding "2" to the sections. This will add additional columns to the report – there will be a set of columns added for every section you add using the "Add" button. The

names will be altered to indicate which section they belong to (*2*pubcount, wghtd_*2*pubcount, *2*pubcount_pos1, wghtd_*2*pubcount_pos1, etc.).

The **Publications** report contains one row for each colleague's publications. Each colleague is identified by the unique identifier Setnb. There is one row in this report per each colleague's publication.

Field	Type	Description
setnb **(key)**	Text	Star unique identifier
pmid **(key)**	Number	Unique article identifier
Journal_name	Text	Name of journal
Year	Number	Year of publication
Month	text	Month of publication
Day	Number	Day of publication
Title	Text	Article title
Volume	Text	volume number of the journal in which the article was published
Issue	Text	Issue in which the article was published
Position	Number	Position in authorship list for the colleague
Nbauthors	Number	Number of coauthors (including star)
Bin	Number	From I to IV
Pages	Text	Page numbers
grant_id	Text	Grant number
grant_agency	Text	Agency who awarded the grant
publication_type	Text	Publication Type from PubMed

The **Star Colleagues report** contains a set of rows for each colleague and star. Each of these sets of rows consists of one row for each year that the star and colleague coauthored at least one paper together. This report will exclude any line for which there are no publications in common for the star and colleague for that year (i.e. nbcoauth1 = 0). So if a star and colleague only coauthored in 1976 and 1984, there will be two rows in this report for them.

The report is grouped by star, colleague and year, with various aggregations performed on the colleague's publications for that year. If the same colleague is a colleague of two different stars, then there will be two different groups in the report for that colleague, one for the first star and one for the second star.

A journal weights file must be provided in order to calculate the weighted publication counts – the software must prompt the user for the location of this file before the reports are run.

Field	Type	Description
setnb **(key)**	Text	Colleague unique identifier
star_setnb **(key)**	Text	Star colleague unique identifier
year **(key)**	Number	Year of publication
Nbcoauth1	Number	Total number of coauthorships (any pos to any pos)
Wghtd_Nbcoauth1	Number	Weighted number of coauthorships (any pos to any pos)
Nbcoauth2	Number	Total number of coauthorships (either star or colleague 1st or last)
Wghtd_Nbcoauth2	Number	Weighted number of coauthorships (either star or colleague 1st or last)
Nbcoauth_1L	number	Number of times the colleague appears as first author on a paper where the star was last author that year
Wghtd_Nbcoauth_1L	number	Weighted number of times the colleague appears as first author on a paper where the star was last author that year
Nbcoauth_L1	number	Number of times the colleague appears as last author on a paper where the star was first author that year
Wghtd_Nbcoauth_L1	number	Weighted number of times the colleague appears as last author on a

		paper where the star was first author that year
Nbcoauth_1M	number	Number of times the colleague appears as first author on a paper where the star was in the middle that year
Wghtd_Nbcoauth_1M	number	Weighted number of times the colleague appears as first author on a paper where the star was in the middle that year
Nbcoauth_M1	number	Number of times the colleague appears as in the middle on a paper where the star was first author that year
Wghtd_Nbcoauth_M1	number	Weighted number of times the colleague appears as in the middle on a paper where the star was first author that year
Nbcoauth_MM	number	Number of times the colleague appears as in the middle on a paper where the star was in the middle that year
Wghtd_Nbcoauth_MM	number	Weighted number of times the colleague appears as in the middle on a paper where the star was in the middle that year
Nbcoauth_LM	number	Number of times the colleague appears as last author on a paper where the star was in the middle that year
Wghtd_Nbcoauth_LM	number	Weighted number of times the colleague appears as last author on a paper where the star was in the middle that year
Nbcoauth_ML	number	Number of times the colleague appears as in the middle on a paper where the star was last author that year
Wghtd_Nbcoauth_ML	number	Weighted number of times the colleague appears as in the middle on a paper where the star was last author that year
Nbcoauth_21	number	Number of times the colleague appears as second author on a paper where the star was first author that year
Wghtd_Nbcoauth_21	number	Weighted number of times the colleague appears as second author on a paper where the star was first author that year
Nbcoauth_12	number	Number of times the colleague appears as first author on a paper where the star was second author that year
Wghtd_Nbcoauth_12	number	Weighted number of times the colleague appears as first author on a paper where the star was second author that year
Nbcoauth_2M	number	Number of times the colleague appears as second author on a paper where the star was in the middle, not NTL that year
Wghtd_Nbcoauth_2M	number	Weighted number of times the colleague appears as second author on a paper where the star was in the middle, not NTL that year
Nbcoauth_M2	number	Number of times the colleague appears as middle author, not NTL on a paper where the star was second author that year
Wghtd_Nbcoauth_M2	number	Weighted number of times the colleague appears as middle author, not NTL on a paper where the star was second author that year
Nbcoauth_2L	number	Number of times the colleague appears as second author on a paper where the star was last author that year
Wghtd_Nbcoauth_2L	number	Weighted number of times the colleague appears as second author on a paper where the star was last author that year
Nbcoauth_L2	number	Number of times the colleague appears as last author on a paper where the star was second author that year
Wghtd_Nbcoauth_L2	number	Weighted number of times the colleague appears as last author on a paper where the star was second author that year
Nbcoauth_2NTL	number	Number of times the colleague appears as second author on a paper where the star was next-to-last author that year
Wghtd_Nbcoauth_2NTL	number	Weighted number of times the colleague appears as second author on a paper where the star was next-to-last author that year
Nbcoauth_NTL2	number	Number of times the colleague appears as next-to-last author on a paper where the star was second author that year
Wghtd_Nbcoauth_NTL2	number	Weighted number of times the colleague appears as next-to-last author on a paper where the star was second author that year
Nbcoauth_NTL1	number	Number of times the colleague appears as next-to-last author on a paper where the star was first author that year
Wghtd_Nbcoauth_NTL1	number	Weighted number of times the colleague appears as next-to-last author on a paper where the star was first author that year

Nbcoauth_1NTL	number	Number of times the colleague appears as first author on a paper where the star was next-to-last author that year
Wghtd_Nbcoauth_1NTL	number	Weighted number of times the colleague appears as first author on a paper where the star was next-to-last author that year
Nbcoauth_NTLM	number	Number of times the colleague appears as next-to-last author on a paper where the star was in the middle, not 2nd author that year
Wghtd_Nbcoauth_NTLM	number	Weighted number of times the colleague appears as next-to-last author on a paper where the star was in the middle, not 2nd author that year
Nbcoauth_MNTL	number	Number of times the colleague appears as middle author, not 2nd author on a paper where the star was next-to-last author that year
Wghtd_Nbcoauth_MNTL	number	Weighted number of times the colleague appears as middle author, not 2nd author on a paper where the star was next-to-last author that year
Nbcoauth_NTLL	number	Number of times the colleague appears as next-to-last author on a paper where the star was last author that year
Wghtd_Nbcoauth_NTLL	number	Weighted number of times the colleague appears as next-to-last author on a paper where the star was last author that year
Nbcoauth_LNTL	number	Number of times the colleague appears as last author on a paper where the star was next-to-last author that year
Wghtd_Nbcoauth_LNTL	number	Weighted number of times the colleague appears as last author on a paper where the star was next-to-last author that year
Frst_collab_year	Number	Year of first collaboration with the star (any position to any position)
Last_collab_year	Number	Year of last collaboration with the star (any position to any position)

4 Generating the Social Networking Report

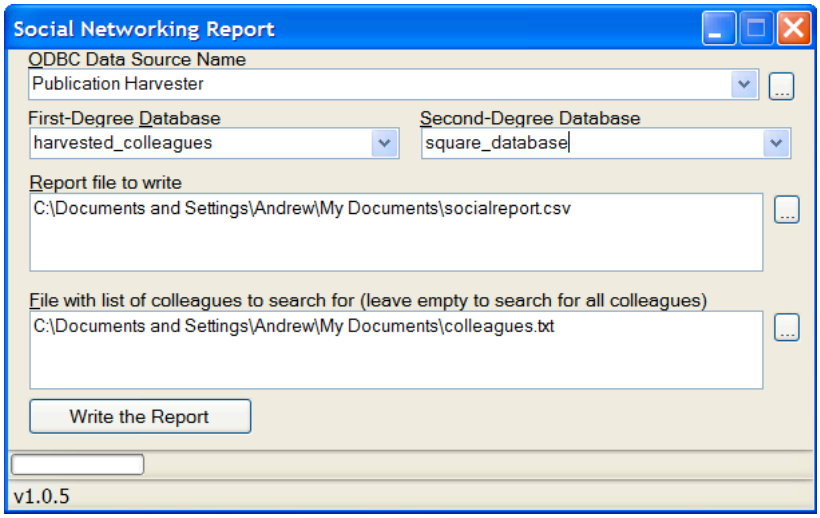
The purpose of the Social Networking Report is to generate a data that will be used to statistically test social networking hypotheses using the data gathered the Publication Harvester and SC/Gen. It creates a two-degree social network: each colleague is connected to several stars, and each star is connected to additional stars.

For additional information on the social networking, including how the social network is defined, see the Social Networking requirements: http://stellman-greene.com/SCGen/SRS_SocialNetworking.doc – it also defines the formats of the report file and the colleagues file.

You can use the Social Networking program to generate this second-degree network:

1. Specify the regular and square databases that have already been generated using Publication Harvester and SC/Gen.
2. By default, the program generates the complete social network for every colleague in the database. But you can specify a list of colleagues to make the software generate a subset of the data.
3. The program searches the database and creates a report file that can be used for statistical analysis.

The Social Networking Report needs two databases: a database that contains the colleagues and defines the first-degree network, and a database that defines the second-degree network. Both databases must be on the same server. To specify the databases, first select an ODBC data source:



Once the data source is selected, the First-Degree and Second-Degree Database dropdown lists are populated with all of the databases on the server. Select the two databases. After the databases are selected, click the "…" button next to the report file box to specify the name of the report file to write. Click the "Write the Report" button to write the report to the file.

You can optionally specify a file that contains a list of setnb values of the colleagues to include in the report. If this file isn't specified, all of the colleagues will be included in the report. If it is specified, then only the colleagues whose setnbs appear in the file will be included in the report.

4.1 Before you can generate the Social Networking report

The Social Networking report generator requires **two** databases that reside on the same server. The first database is a standard database that has been generated by the Publication Harvester and SC/Gen, with all of the colleagues generated. If you have followed all of the steps in this manual through the end of section 3, then you have created the first database.

The second database is also created by the Publication Harvester and SC/Gen, using the same Publication Harvester input file. The difference is that for the first-degree database you'll use a **square roster file** when you use SC/Gen to generate the colleagues, while you'll use a larger roster to generate the second-degree database. Typically, your roster file for the second-degree database will be much larger than the input file for the Publication Harvester, because you usually want to cast a wide net to find as many colleagues as possible for each person in the Publication Harvester input file to generate a large second-degree social network. However, the first degree social network is created by **starting with the colleagues in the first database**. Each of those colleagues is associated with a person in the original input file. (Some are associated with more than one person; in that case, they're treated as two separate cases and have two separate sets of rows in the social networking report.) By feeding a square database to SC/Gen when you generate the first-degree database, you create a smaller list of potential colleagues.

Once all of the people are found for each colleague, then ***their*** colleagues need to be found. But those colleagues aren't found in the original database. Instead, the program searches for second degree colleagues ***using the people in the original input file*** that was used to create the Publication Harvester database.

To do this, you'll create a **square roster**, which is just a roster that contains the people in the original input file. You'll use the Publication Harvester to create a database with the same input file you used for the regular database. Then you'll use SC/Gen to find the colleagues in the square database, using the square roster file instead of the one you used for the regular database.

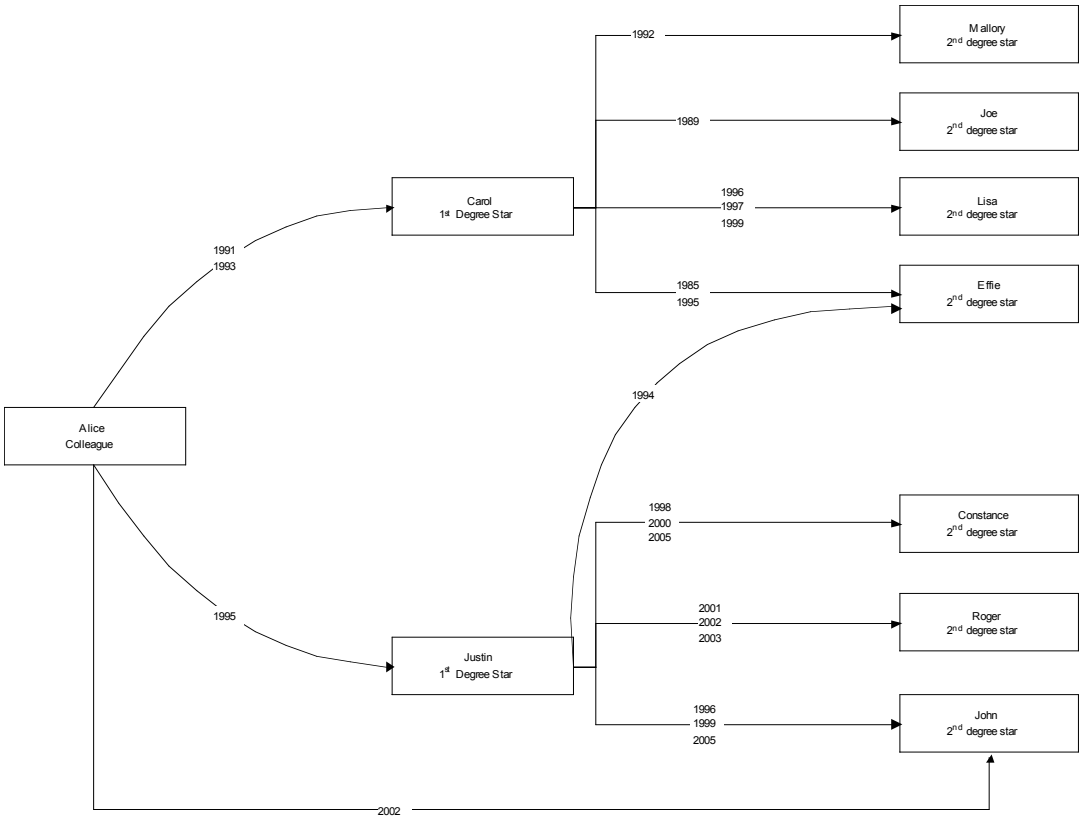
Note: You don't have to re-harvest the publications to create the square database. You can just copy the MySQL database: http://dev.mysql.com/doc/refman/5.0/en/upgrading-to-arch.html

Once you have the two databases (the regular database and the square database), you're ready to use the Social Networking report software to generate the report.

4.2 The second-degree social network

The purpose of the Social Networking Report is to generate a report that will be used to statistically test social networking hypotheses using the data gathered the Publication Harvester and SC/Gen. It generates a second-degree social network using *two* databases generated by Publication Harvester and SC/Gen. The second-degree database is a normal database generated by SC/Gen, while the first-degree database should be a "square" database that was generated by harvesting the same people that were used to harvest the first database. The difference is that instead of using the full roster of potential colleagues, its roster only contains the people in the original people list – that way, it will only find colleagues who are also on the list.

Once the two databases are created, the social network can be found for each colleague. Here's a typical colleague's social network:



This picture shows the social network for a colleague named Alice, who is the colleague of two people on the original list: Carol and Justin. Those people appeared in the list that was fed into Publication Harvester, and when their publications were downloaded Alice appeared on both of them. She coauthored publications in 1991 and 1993 with Carol, and in 1995 with Justin.

Mallory, Joe, Lisa, Effie, Roger and John also appeared on the list that was fed into Publication Harvester. A second database contains the "square" data – the colleague lists for each person restricted to only the people who were in the original list. Since Alice is a colleague of Carol, who is a colleague of Mallory, then Mallory is a second-degree colleague of Alice.

4.3 Generating the social network report

The goal of the report is to identify the way the relationship between Alice and Bob changes on a year-by-year basis. To do this, it contains a set of rows for each combination of colleague, first degree star and second degree star. In this example, there are four combinations (Alice, Carol, Bob; Alice, Eve, Bob; Alice, Justin, Bob; Alice, Mallory, Bob). Each of these four sets of rows will contain one row per each year in the range from the earliest publication to the latest publication. For example, the earliest publication in the Alice-Carol-Bob combination is 1991, and the latest is 2003, so for these three people there will be 13 rows in the report.

For each row, there are three relationships to measure: colleague and 1st degree star; 1st degree star and 2nd degree star; colleague and 2nd degree star. This last measurement is important in order to answer research questions about whether being in the same network as a star would predict a 2nd degree colleague's work. It also may be useful to limit a network only to colleagues who never end up working with 2nd degree stars, or to years that predate colleague-2nd degree star coauthorships.

Each of the three relationships in a given year is measured using stock and flow. The flow measurement is the number of publications in coauthored by the two people in that year. The stock measurement is the cumulative number of publications coauthored by the two people in that year or any earlier year.

The report will be a comma-delimited text (CSV) file with a header row and the following columns:

Field	Type	Description
setnb0 **(key)**	Text	Colleague unique identifier
setnb1 **(key)**	Text	1st degree star colleague unique identifier
setnb2 **(key)**	Text	2nd degree star colleague unique identifier
year **(key)**	Number	Year
flow0to1	Number	Flow for colleague to 1st degree star
stk0to1	Number	Stock for colleague to 1st degree star
flow1to2	Number	Flow for 1st degree star to 2nd degree star
stk1to2	Number	Stock for 1st degree star to 2nd degree star
flow0to2	Number	Flow for colleague to 2nd degree star
stk0to2	Number	Stock for colleague to 2nd degree star

In the social network shown in the above diagram, there are 14 different combinations of colleague, 1st degree star and 2nd degree star in this network. The following rows will be generated for the Alice-Justin-John combination:

setnb0	setnb1	setnb2	year	flow0to1	stk0to1	flow1to2	stk1to2	flow0to2	stk0to2
Alice	Justin	John	1995	1	1	0	0	0	0
Alice	Justin	John	1996	1	1	1	1	0	0
Alice	Justin	John	1997	1	1	0	1	0	0
Alice	Justin	John	1998	1	1	0	1	0	0
Alice	Justin	John	1999	1	1	1	2	0	0
Alice	Justin	John	2000	1	1	0	2	0	0
Alice	Justin	John	2001	1	1	0	2	0	0
Alice	Justin	John	2002	1	1	0	2	1	1
Alice	Justin	John	2003	1	1	0	2	1	1
Alice	Justin	John	2004	1	1	0	2	1	1
Alice	Justin	John	2005	1	1	1	3	1	1

The following rows will be generated for the Alice-Carol-Effie combination:

setnb0	setnb1	setnb2	year	flow0to1	stk0to1	flow1to2	stk1to2	flow0to2	stk0to2
Alice	Carol	Effie	1985	0	0	1	1	0	0
Alice	Carol	Effie	1986	0	0	1	0	0	0
Alice	Carol	Effie	1987	0	0	1	0	0	0
Alice	Carol	Effie	1988	0	0	1	0	0	0
Alice	Carol	Effie	1989	0	0	1	0	0	0
Alice	Carol	Effie	1990	0	0	1	0	0	0
Alice	Carol	Effie	1991	1	1	1	0	0	0
Alice	Carol	Effie	1992	0	1	1	0	0	0
Alice	Carol	Effie	1993	1	2	1	0	0	0
Alice	Carol	Effie	1994	0	2	1	0	0	0
Alice	Carol	Effie	1995	0	2	1	2	0	0

4.4 Restricting the Report to a List of Colleagues

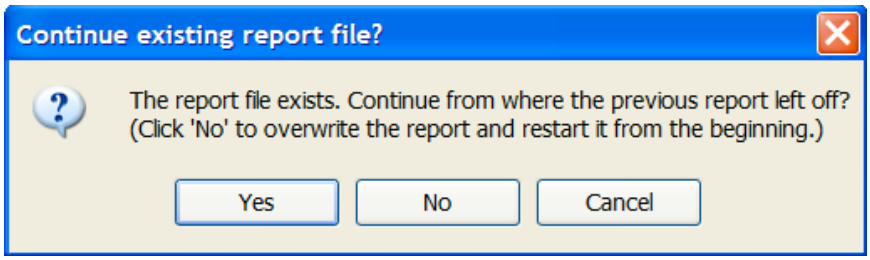
A SQL database generated by the Publication Harvester and SC/Gen can contain a very large number of colleagues – and most of those colleagues may be irrelevant to the researcher generating the reports. To remedy this, the researcher may optionally specify a list of colleagues. If this list is specified, then the software only generates the social network for those colleagues, and the final report contains only those colleagues. The list of colleagues is text file that contains the setnb values for each colleague to include in the report, with one setnb per line. For example, if you want to generate a report that only includes Alice, Justin and Efffie, you'd specify a text file with their setnbs on each line:

```
Alice
Justin
Effie
```

If this file is specified, then the software will only generate a report for Alice, Justin and Effie. However, the software will not generate any report rows for any setnb that is not actually contained in the database.

4.5 Fault tolerance

The Social Networking report generator is capable of producing extremely long reports, which may take a long time to produce. If a long run were to be interrupted (by a power failure, for example), it would be frustrating and time-consuming to have to regenerate the report rows that were already produced. To avoid this problem, fault tolerance is built into the report generator. To resume a partially generated report, just select the report file that you began. The software will first warn you that you have selected an existing file. It will then display a prompt:



If you select "Yes", the report generator will rename the existing file, adding ".bak" to the end of the filename. If will then create a new file and copy all of the rows out of the old one – except for the block of rows for the last colleague in the file, because if the report generator was previously interrupted, it may not have finished writing that colleague's network. It will then skip any colleague which was copied from existing file.

If you select "No", the report generator will delete the existing file. If you select "Cancel", then it will cancel the operation and not write or change any files.

GNU Free Documentation License

Version 1.2, November 2002

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover

must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- **A.** Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- **B.** List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- **C.** State on the Title page the name of the publisher of the Modified Version, as the publisher.
- **D.** Preserve all the copyright notices of the Document.
- **E.** Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- **F.** Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- **G.** Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- **H.** Include an unaltered copy of this License.
- **I.** Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- **J.** Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- **K.** For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- **L.** Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- **M.** Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- **N.** Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- **O.** Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

5 Revision History

Date	Author	Description
19-Dec-2007	Andrew Stellman	Created initial version
10-Jan-2008	Andrew Stellman	Fixed screenshots, filled out overview and compelted social network report section
11-Jan-2008	Andrew Stellman	Added "Before you generate the social networking report" section, fixed screenshots (SC/Gen had a typo in a label)
23-Jan-2008	Andrew Stellman	Added fault tolerance to the Social Networking report, and changed "square" and "regular" databases to "first-degree" and "second-degree".