

MEDICI. Program Structure and User Guide

In the following, the user interface screens are shown with an explanation of each, with a mention of any related functionality between the different screens.

The source code written in Java/Java FX and system runtime is publicly available in the Github link "<https://github.com/dnettlet/MEDICI>". *The system requires Eclipse/Java runtime (or similar) to be installed in your computer.*

As a first test run, it is recommended the user runs the system in default mode with completely pre-assigned settings and example graph and community files. Take a look at where the input, default and output files are located with the file explorer. Then go directly to the "generate data" tab, click the blue colored "generate data" button when it completes go to the "results" tab to see the statistics of the output data. This will enable the user to become familiar with the system, before progressing to modifying the profile to community assignments, the profile definitions, and finally the input graph and communities.

As a general vision, Figure 1 shows the dependencies between different screens – while the majority are independent (they are managed directly by rootLayout), some have inter-dependencies. For example, the attribute edit screen depends on the attribute screen, and "Prof. Comm Assign" depends on "Comm. & Profiles".

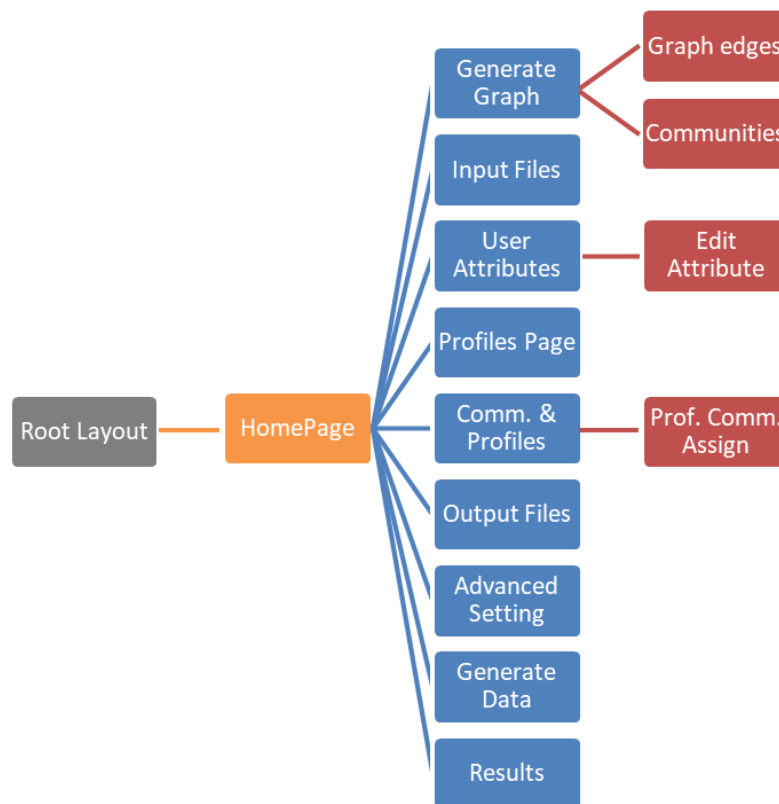


Fig. 1. Hierarchy of interface screens

A.1 Root Layout

The Root Layout is not a screen as such, but the basis for all the user interface screens, and contains just a bar in the upper menu. It is always visible given that it allows fast access to key functions.

It has a basic structure with different pull-down menus which offer the different options (see Figs. 2). The first menu is “File” (Fig. 2a), which gives access to the functionality of “Import Config” to import a configuration from a file; “Export Config” to export the actual configuration into a file; “Close” to finalize the execution and close the program.



Fig. 2a Main menu pulldown list (File)



Fig. 2b Main menu pulldown list (Switcher)

The second menu, called “Options”, gives direct access to any of the screens in the user interface.



Fig. 2c Main menu pulldown list (Help)

Lastly, the “Help” menu has the following options: “Documentation” opens the user manual of the program as a PDF file; “About” opens a pop-up window with the credits which show the authors of the program: “License” opens a pop-up window showing a summary of the software license with a button to see the full license details. The software is licensed under the GNU General Public License 3.0.

A.2 Home Page

The first page is shown when the programs starts up (Fig. 3), and just shows a “go” button which then navigates to the first tab, “Generate Graph”. Also, the pull down menus are available, as described before.



Fig.3 Home Page

A.3 Generate Graph and identify Communities

Figure 4 shows the first screen of the program “workflow” (going left to right) which allows the user create the graph from scratch using the RMat graph generator, and then identify the communities using the Louvain community detection algorithm. In the case of RMat, it is recommended to assign the number of nodes and edges, leaving parameters a to d as default. However, experts may try varying these to evaluate the resulting graphs in terms of community structure. Essentially, parameters a and d define the communities whereas b and c define the links between communities. In the case of Louvain, the algorithm has been customized to find exactly 10 communities, and a “quality” value is shown after processing of the assigned communities. The closer the quality value is to 1, the better the quality.

Note that in the upper part of the screen (Figure 4) is the navigation bar with different tabs common to all workflow screens. There is one tab for each main workflow step, allowing direct access, with the current one being highlighted.

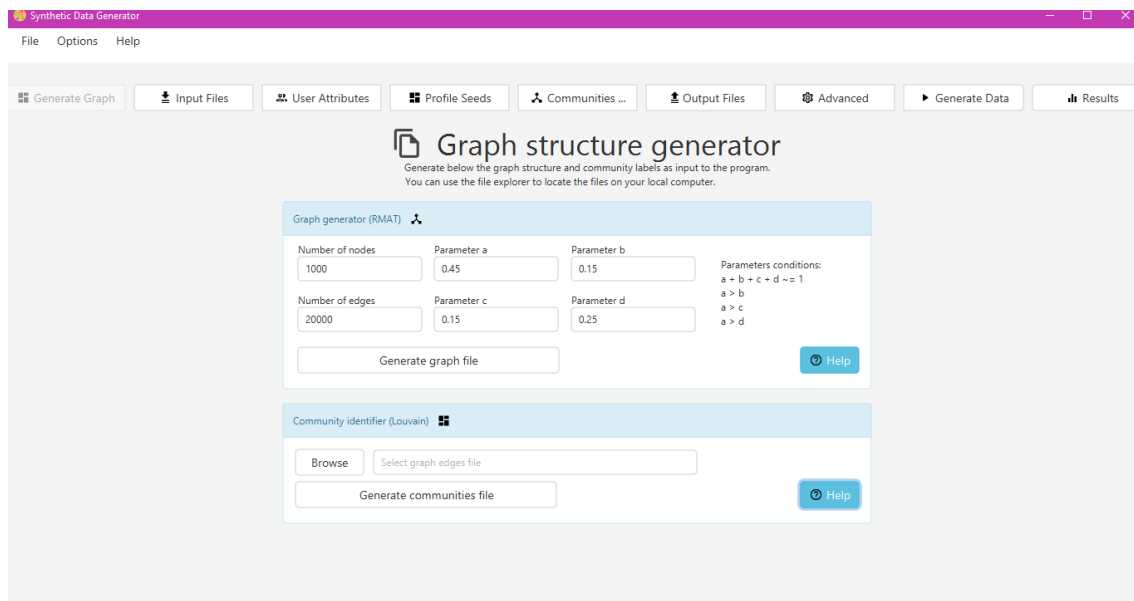


Fig. 4 Graph Generator Page

The first file generated is the graph structure, which is a simple list of node pair connections in “txt” format. For example:

1kby30k.csv

12	77
99	697
843	847
316	66
845	331
557	278
734	736
750	752
166	175
507	517
694	688
328	328

where in the first line 12;77 defines two nodes (or “users”) 12 and 77 and the link between them.

The second input file contains the community label assignments for each node:

1kby30kcommunities.csv

0	0
1	1
2	2
3	1
4	0
5	0
6	1
7	2

where in the first line *0;0* indicates that node 0 is assigned to community 0, the second line *1;1* indicates that node 1 is assigned to community 1, and so on.

Note that the default files are in “./resources/Default_files” and if the user generates new graph and community files they will appear in the folder “./resources/Input_files”.

A.4 Input File Settings

Figure 5 shows the second screen of the program “workflow” which allows the user to select the input files for the graph and the communities, which will be used later by the “generate data” process. The default files are in the system folder “./resources/Default_files”. file names will be automatically assigned (in a default system folder which is shown). If the user generates new graph and community files using the RMAT and Louvain algorithms, they will appear in the folder “./resources/Input_files”. Also, the user can place his/her own graph and community files in the folder “./resources/Input_files”.

The red button allows the user to reset the file names and paths to default, the light blue button displays a popup window with additional information and the green button saves the current assignment and go to the next step in the workflow.

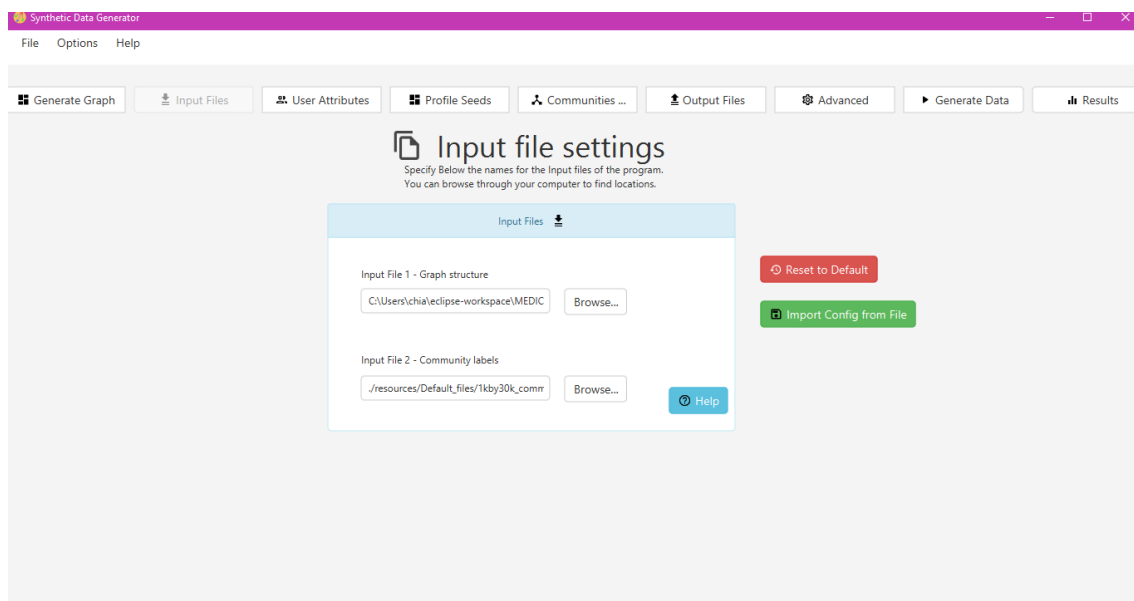


Fig. 5 Input File Settings Page

A.5 User Attributes

Figure 6 shows the “User Attributes” page which allows the user to visualize the attributes which characterize the individuals in the social network, and modify the relative proportions of each attribute-value. Each attribute is presented as a record with name, description and a list of values with their respective frequencies. For example, on the left, attribute “Age” has seven ranges, the first being 18-25 years with proportion 0.25 and the second 26-35 years with proportion 0.25. In contrast, age

range 76-85 years has the relatively smallest proportion of 0.08. Thus the distribution defined for “age” has a strong bias towards younger people.

On clicking “edit” a window will appear to edit the attribute (see next Section A6 and Fig. 7). Note the sum of the proportions for the values of given attribute must sum to 1.0.

In terms of the data generation/propagation, these percentage distributions and categories are used when assigning neighbor and non-neighbors of the seed (profile) nodes, which has a probabilistic nature.

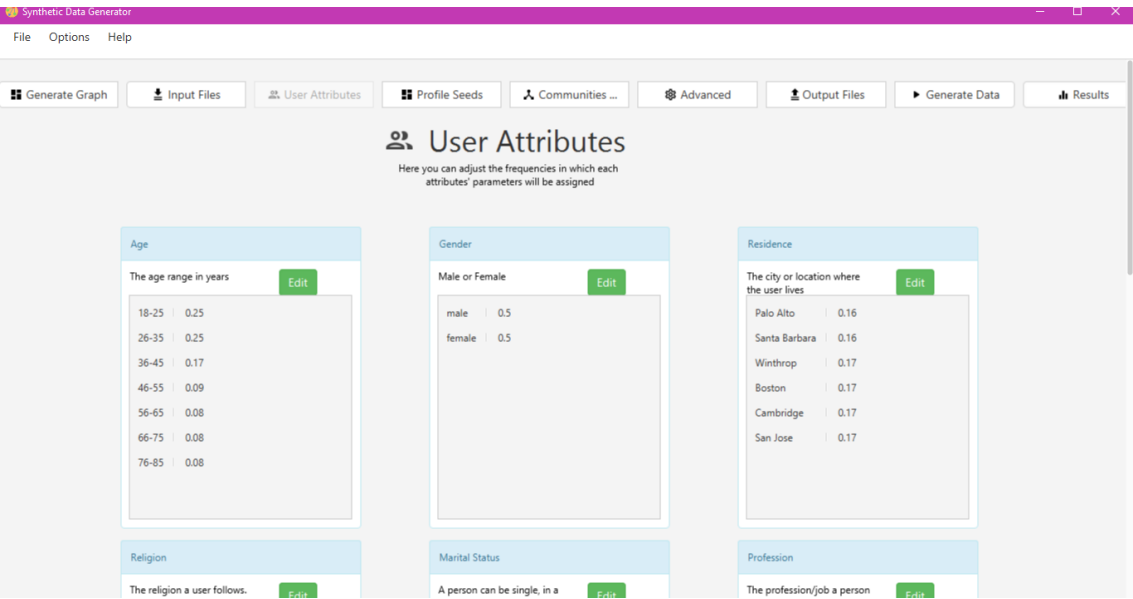


Fig. 6 User Attributes Page

A.6 Attribute Edit Dialog

Fig. 7 shows the window which allows the user to edit the attribute-value proportions. The attribute record shows the attribute name and description, which is followed by a list of possible values for the attribute and the relative frequency/proportion for each values.

For example, in Fig. 7, the attribute “gender” has been assigned two possible values, “male” and “female”, which currently have assigned equal relative proportions of 50% each. If a modification is made, the green “ok” button will save the changes or click the red “cancel” button to quit without saving.

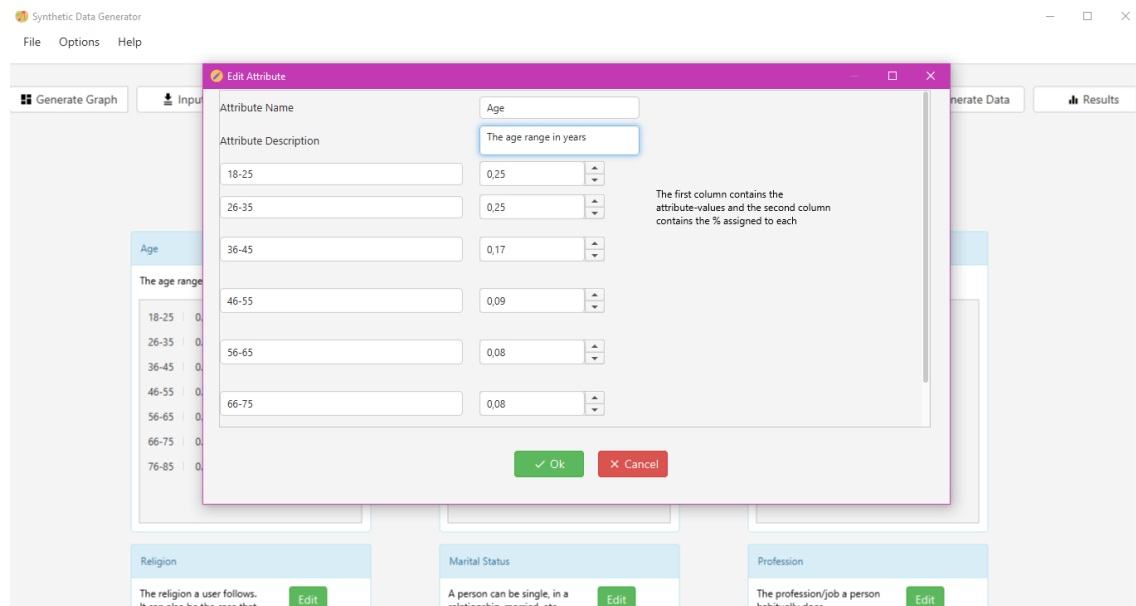


Fig. 7 Attribute Edit Dialog

A.7 Profiles Seeds Page

Figure 8 shows the “Profile Seed Settings” page. In this screen, different user profiles can be configured. One profile is later assigned to each community (see Section A8), which are fixed to 10 profiles/communities. For each profile there is a pulldown list for each attribute, with the assignable values. There is also a (red) button on the right to return all the fields to their default values. For example, in Fig. 8, profile 0, attribute “age” is set to the range “36-45” and the “marital status” attribute is set to “Married”.

The user can customize a profile by choosing an “attribute-value” for each of the attributes which describe a “seed” (or prototype) individual of the social network. The system allows for up to 10 profiles to be created, one for each community. When the data is generated the Medici algorithm probabilistically assigns the seeds “neighbors” with profiles which are similar to the seed, with a controlled amount of random noise.

Profile Seed Settings
Define profiles: assign attribute-values to profile seeds.

Profile 0

36-45	No religious affiliation	Center	music artist
male	Married	Heterosex	drink brand
Boston	Sales and office	tv show	

Profile 1

26-35	Buddhist	Center	tv show
female	Divorced	Heterosex	soccer club
Cambridge	Sales and office	entertain	

Profile 2

18-25	Christian	Center Left	drink brand
male	Single	Heterosex	entertain
Palo Alto	Student	drink brand	

Profile 3

18-25	Muslim	Center Right	tv show
female	Single	Heterosex	soccer club
Winthrop	Professional	music artist	

Profile 4

56-65	Hindu	Right	entertain
-------	-------	-------	-----------

Profile 5

66-75	Jewish	Far Left	tv show
-------	--------	----------	---------

[Reset to Default](#)

Fig. 8 Profile Seed Settings Page

A.8 Communities and Profiles

Once the profile seeds have been defined the next step is to assign them to the communities. The communities have relative proportions depending on how many nodes they have assigned, which is defined when the graph is generated and the community labels assigned. Hence, if you wish to have a majority of users with a given profile P^n in the overall graph, you find the biggest community and assign the profile P^n to it. Contrastly, if you want a just a few users to have a given profile, you find a community with a small proportion and assign the profile to it.

Profile to Community assignment is accessed using the screen shown in Fig. 9. This screen has a button which opens the window to assign the profiles to communities, and includes informational text explaining the functionality.

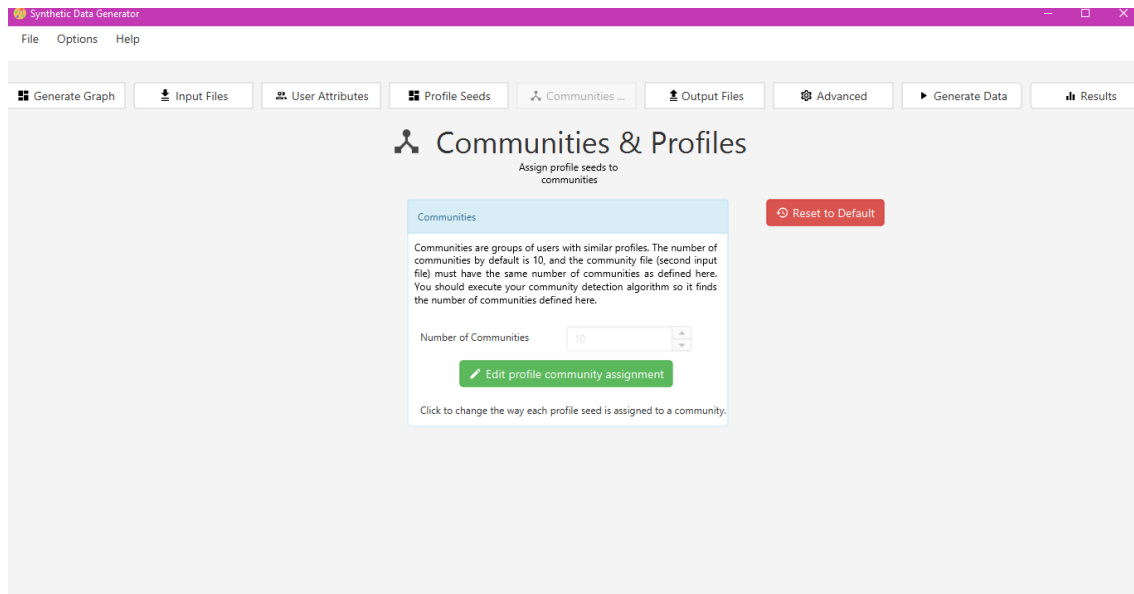


Fig. 9 Communities & Profiles page

Click on the green button labelled “Edit profile community assignment” (Fig. 9) to open the screen (Fig. 10) which allows the user to see the frequency/proportion for each community and assign the profiles. The total (sum for all profiles) is shown which must sum to 100% (this is calculated automatically from the graph). For example, profile 0 has a proportion of 21.6%, and profile 7 has just 0.5%. It can be seen that profiles 0,1,2 and 5 have the highest proportions and profiles 7 and 9 have the lowest.

Click on the green button labelled “Edit profile community assignment” (Fig. 9 right) to open the screen which allows the user to assign the profiles to communities (Fig. 10b). In this screen there is a list of communities and for each community a pulldown list where there user can assign the profile. It can be seen that profile 0 has been assigned to community 0 and profile 6 to community 8. The graph generator and community labeller produce a distribution of proportions similar to a “long tail”, within the limitations offered by 10 values, thus having a similar community size distribution to real world social networks.

Note that the user can provide his/her own graph and community files, instead of generating them in Medici, but then the responsibility for their correctness relies on the user.

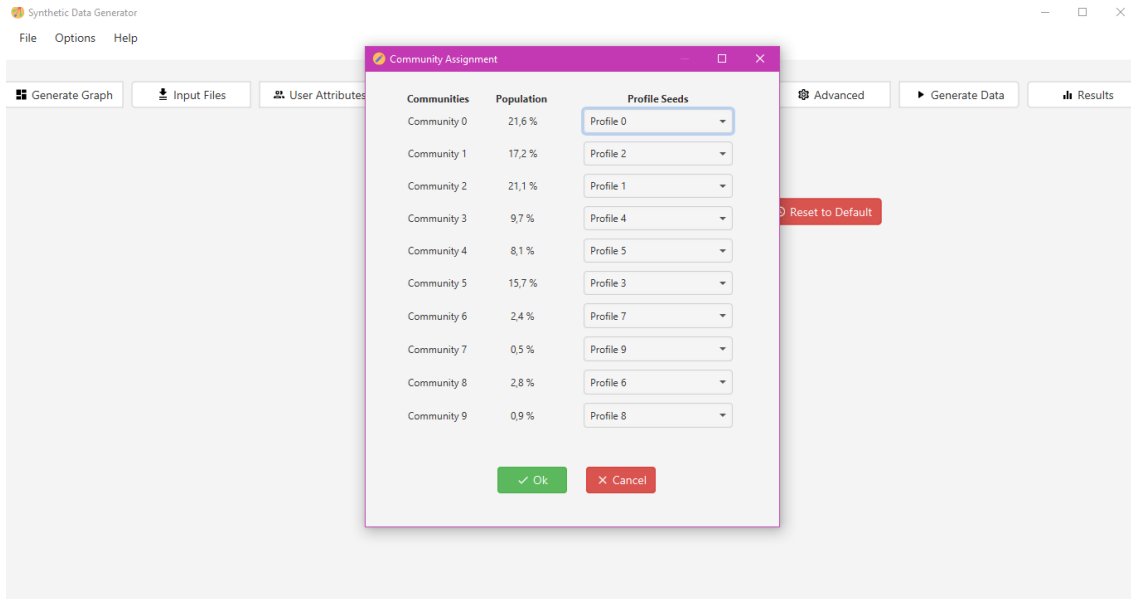


Fig. 10 Assignment of Profiles to Communities

A.9 Output Files

Figure 11 shows the screen to assign the output files. It is similar to the input file screen (Fig. 5), with a similar functionality. It allows the user to specify where to save and what to call the files which contain the results of generating the social network user data. The default folder is “./resources/Output_files/”. The user can also save the current configuration (attribute and profile assignments, etc.) by clicking the green button “Export Config to File” on the left. The red button restored the default system values.

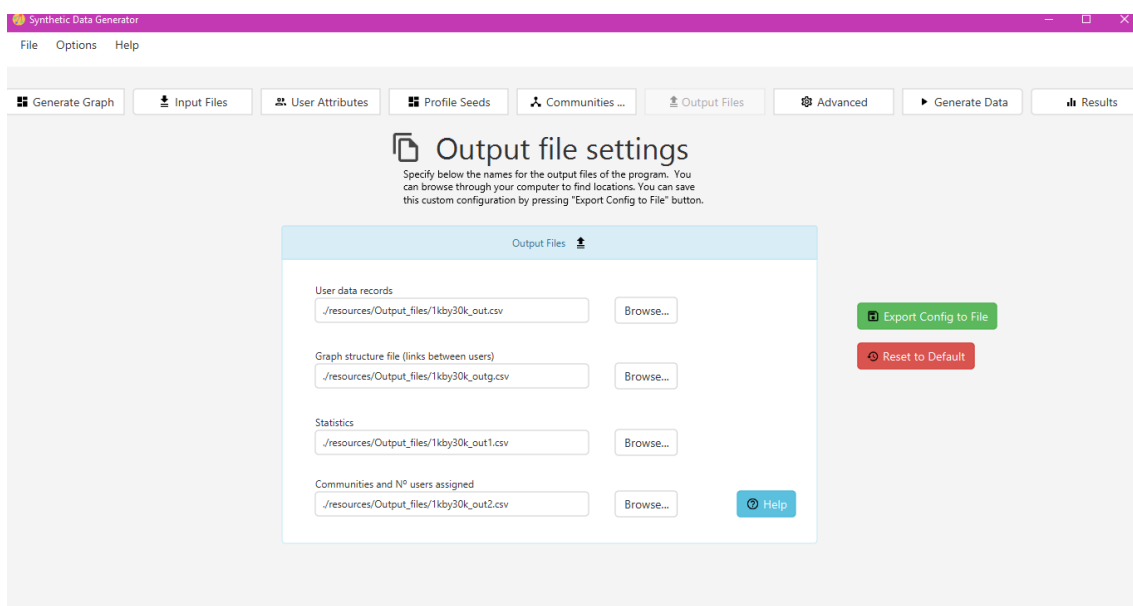


Fig. 11 Output File Settings

The format of each output file is as follows:

User data records - default name “1kby30k_out.csv”

user	age	gender	residence	religion	maritalstatus	profession	politicalorientation	sexualorientation	numfriends	like1	like2	like3	classvalue	auth	community
999	26-35	female	Winthrop	Hindu	Divorced	Professiona	Left	Heterosexual	LOW	entertainment	entertainment	music artist	NO	0.026178	5
998	26-35	female	Cambridge	Buddhist	Single	Service	Center Right	Heterosexual	LOW	entertainment	entertainment	entertainment	YES	0.041885	6
997	26-35	female	Boston	Jewish	Married	Service	Left	Heterosexual	LOW	music artist	music artist	entertainment	NO	0.031414	3
996	18-25	female	Winthrop	Muslim	Married	Professiona	Center Right	Heterosexual	LOW	music artist	music artist	drink brand	NO	0.068063	5
995	66-75	female	San Jose	Jewish	Married	Production	Far Left	Heterosexual	LOW	entertainment	tv show	entertainment	NO	0.036649	4
994	18-25	female	Winthrop	Muslim	Single	Production	Center Right	Heterosexual	LOW	drink brand	tv show	drink brand	NO	0.073298	5
993	26-35	female	Cambridge	Buddhist	Widowed	Sales and o	Center Left	Homosexual	LOW	entertainment	tv show	soccer club	YES	0.031414	2
992	26-35	female	Cambridge	Christian	Widowed	Sales and o	Left	Heterosexual	LOW	music artist	music artist	entertainment	YES	0.115183	8
991	56-65	female	Palo Alto	Jewish	Married	Production	Left	Bisexual	LOW	music artist	music artist	entertainment	NO	0.041885	4
990	56-65	male	Santa Barbi	Hindu	Widowed	Natural resi	Left	Heterosexual	LOW	tv show	entertainment	drink brand	NO	0.120419	9
989	56-65	male	Boston	Buddhist	Single	Sales and o	Center Left	Heterosexual	LOW	music artist	music artist	drink brand	NO	0.08377	0
988	26-35	female	Cambridge	Hindu	Married	Student	Left	Heterosexual	LOW	drink brand	tv show	drink brand	YES	0.078534	2
987	18-25	female	Winthrop	Muslim	Single	Professiona	Center Right	Heterosexual	LOW	music artist	tv show	soccer club	NO	0.068063	5
986	56-65	male	San Jose	Hindu	Single	Sales and o	Center Right	Homosexual	LOW	tv show	drink brand	soccer club	NO	0.125654	3
985	26-35	female	Cambridge	Buddhist	Married	Professiona	Left	Heterosexual	LOW	entertainment	tv show	soccer club	YES	0.115183	2
984	26-35	female	San Jose	Sikh	Divorced	Student	Center	Heterosexual	LOW	drink brand	tv show	drink brand	NO	0.198953	2
983	56-65	female	Cambridge	Hindu	Divorced	Production	Far Left	Heterosexual	LOW	entertainment	entertainment	music artist	NO	0.057592	9
982	26-35	female	Cambridge	Buddhist	Single	Sales and o	Center	Heterosexual	LOW	entertainment	tv show	soccer club	YES	0.115183	2
981	56-65	female	San Jose	Hindu	Single	Natural resi	Right	Heterosexual	LOW	music artist	entertainment	entertainment	NO	0.089005	3
980	66-75	female	San Jose	Christian	Single	Production	Far Left	Bisexual	LOW	tv show	drink brand	soccer club	NO	0.141361	4
979	26-35	female	Palo Alto	Buddhist	Single	Manager	Center Left	Homosexual	LOW	music artist	music artist	entertainment	NO	0.078534	0
978	18-25	female	Winthrop	Muslim	Single	Professiona	Center Right	Heterosexual	LOW	music artist	tv show	soccer club	NO	0.136126	5
977	46-55	male	Boston	Hindu	Married	Manager	Center	Heterosexual	LOW	entertainment	entertainment	music artist	NO	0.08377	0
976	18-25	female	Winthrop	Muslim	Single	Professiona	Center Right	Heterosexual	LOW	music artist	tv show	soccer club	NO	0.136126	5
975	18-25	female	San Jose	Buddhist	Widowed	Student	Center	Heterosexual	LOW	drink brand	tv show	drink brand	NO	0.052356	1
974	26-35	female	Cambridge	Hindu	Married	Student	Left	Heterosexual	LOW	entertainment	entertainment	music artist	YES	0.089005	3

Contains the social network user information, one row per individual, where the first row is the header and each consecutive row consists of the user (node) id followed by the value for each attribute for this individual. For example, user id 999 has “gender” attribute assigned with value “female” and has “residence” attribute assigned with value “Winthrop”, etc.

Graph structure file (links between users) - default name “1kby30k_outg.csv”

user	userf	linkweight
999	487	0.57
999	658	0.54
999	428	0.62
999	731	0.30
999	57	0.60
998	448	0.47
998	791	0.89
998	937	0.70
998	299	0.55
998	656	0.56
998	844	0.50

Contains the graph information, one row per link. For example, in the first row user 999 has a link to user 487 with a link strength of 0.57, and the second row shows that user 999 also has a link to user 658 with strength 0.54, etc.

User ids – default name “1kby30k_out1.csv”

community	attribute	value	frequency
ALL	AGE	18-25	275
ALL	AGE	26-35	299
ALL	AGE	36-45	182
ALL	AGE	46-55	46
ALL	AGE	56-65	106
ALL	AGE	66-75	64
ALL	AGE	76-85	28
ALL	GENDER	Male	329
ALL	GENDER	Female	671
ALL	RESIDENCE	PaloAlto	141
ALL	RESIDENCE	SantaBarbara	141
ALL	RESIDENCE	Winthrop	195
ALL	RESIDENCE	Boston	206
ALL	RESIDENCE	Cambridge	198
ALL	RESIDENCE	SanJose	119
ALL	RELIGION	Buddhist	214
ALL	RELIGION	Christian	211
ALL	RELIGION	Hindu	287
ALL	RELIGION	Jewish	87
ALL	RELIGION	Muslim	157
ALL	RELIGION	Sikh	20

Contains the statistics of the data generated for the users, for ALL the graph and then for each community in turn. For example, the first row shows that for ALL the graph, attribute-value “age 18-25” has a frequency of 275.

Communities and N° users assigned – default name “1kby30k_out2.csv”

9	9
8	28
7	5
6	24
5	157
4	81
3	97
2	211
1	172
0	216

Contains the summary information for each community. For example, the first row indicates that 9 users are assigned to community 9 and the last row indicates that 216 users are assigned to community 0.

A.10 Advanced Settings

Figure 12 shows the “Advanced Settings” screen, which displays the options for changing the advanced features of the program. This is only for expert users, and the “thresholds” option is disabled in the current version of Medici. There are three options: “Profile seeds”, “Thresholds” and “Randomness”, which have a brief explanatory text on screen and additional information is available by pressing the light blue “help” button. In the majority of situations the program should function optimally with the default settings of “seeds percentage=11” and “random assignment ratio=low”.

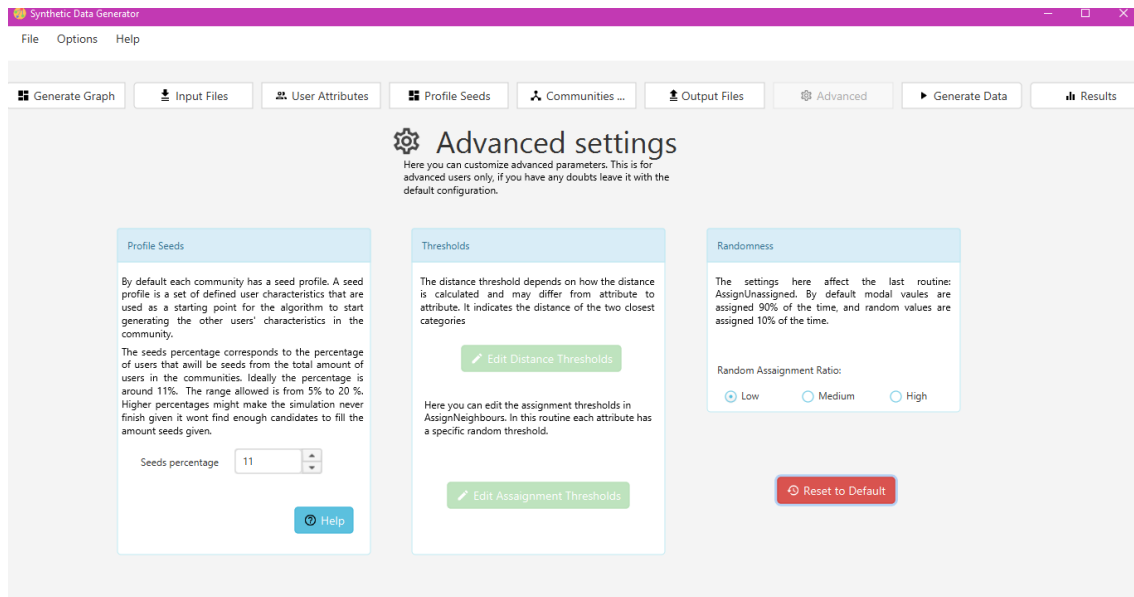


Fig. 12 Advanced Settings Screen

A.11 Generate Data Screen

Figure 13a shows the “Generate Data Screen” where the user finally executes the data generation based on all the attributes, profiles and community assignments which have been assigned previously. The main button “Generate Data” executes the back-end Medici algorithm and while it is running a loading bar appears to show progress (Fig 13b). There is also a red button to cancel the process. These two elements disappear once the data is generated and a pop up window will appear to indicate to the user that the generation has terminated, with a count of how many nodes have been assigned (seeds + neighbours), which in Figure 13b is shown as 786, of the 1000 total nodes in the graph. This means that the remaining 214 nodes will be assigned probabilistically (due to closeness to an assigned node) or randomly.

If any errors have occurred a pop up will appear with a red cross to inform of any problems.

If the process has terminated correctly, the user can now click on the “results” tab to see the statistics of the resulting data and access the output files.

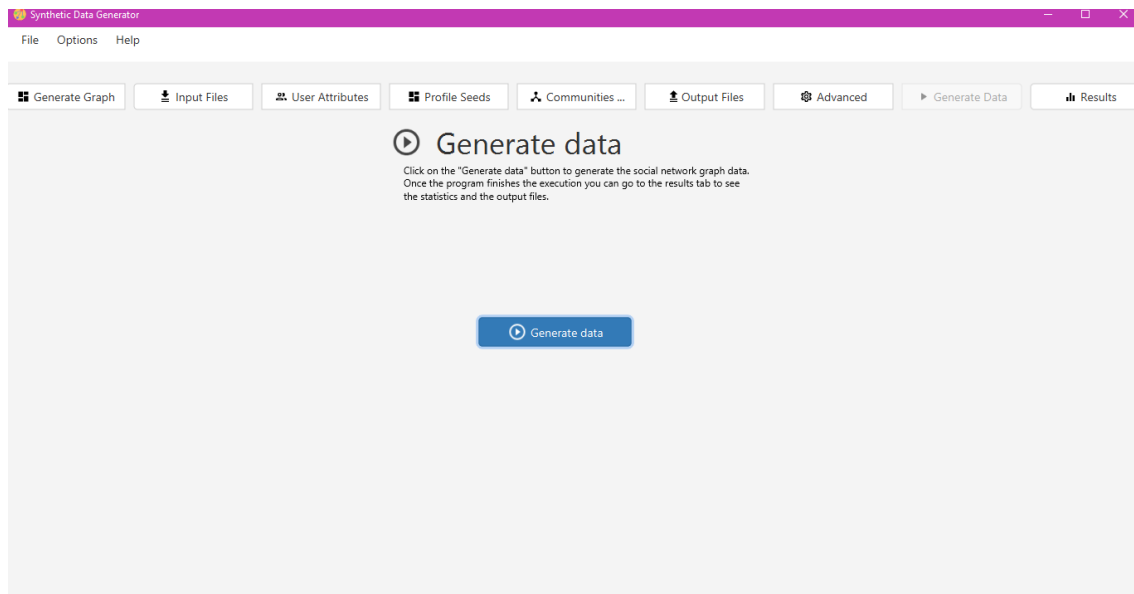


Fig. 13a Generate Data Screen

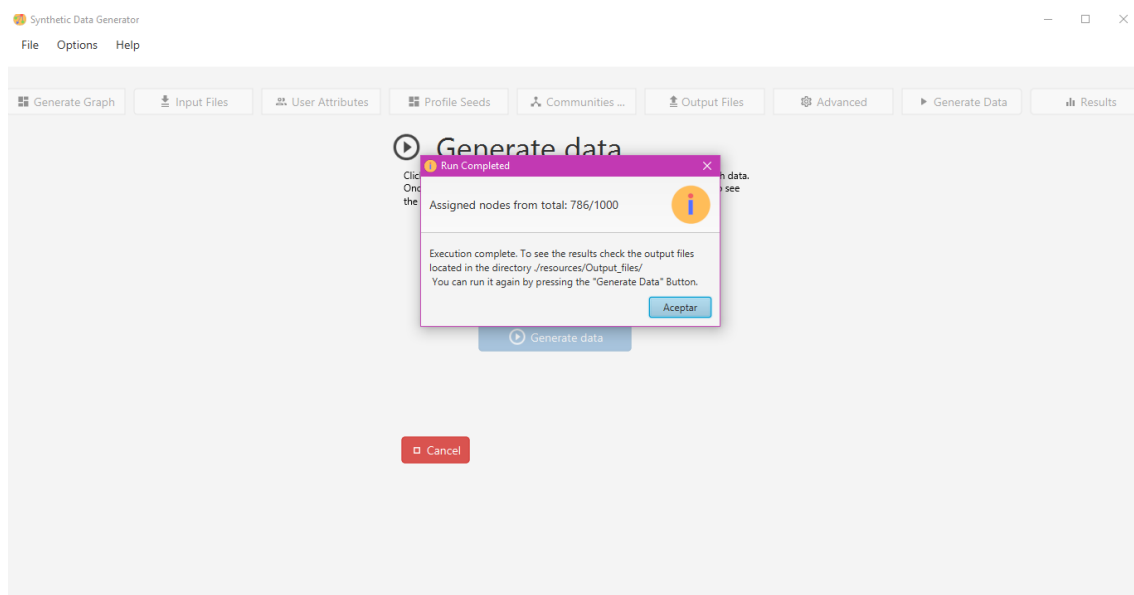


Fig. 13b Run Screen

A.12 Results Screen

Figure 14(a to c) show the results screen with the statistics for the data which has been generated Fig. 14a shows the statistics for the whole graph (all communities), and Figs. 14b and 14c show the statistics just for community 2 (profile1). On the top left of the screen there is an option to select all the data or a specific community. On the top right there are options to open the different results files.

Below there is the statistics panel which can be filtered using the pulldown list on the top right. This allows the user to see the global statistics or filter on a given

community. The statistics show a graphic with the proportion for each community and corresponding profile seed (pie chart on left), then for each attribute a graphic with the percentage of assigned values (e.g. age and gender to the right).

Each attribute is shown as a frequency bar graph with a bar for the profile percentage assigned by the user to the “seeds” and another bar with the frequency for the same attribute in that community.

Here it is important to note that the two percentages will not necessarily be the same, because the algorithm, as described previously, assigns the “pure” profiles only to the seeds in a given community. The neighbors of each seed will be similar but not necessarily the same (e.g. 30% could be different), due to how the propagation works, with a predefined stochastic (random) behavior which guarantees diversity. Furthermore, the remaining nodes (not neighbors of any seed) will have a further diversity. Hence it is normal that the profile frequency and the community frequency will be different most of the time. However, overall, the community should have a distinct identifiable similarity to the assigned profile attribute distributions, which distinguishes it from other communities with different profile assignments.

So in Figure 14a it can be seen that overall (general trends in the graphs) there are three main age groups, 18-25, 26-35 and 36-45, the gender has a bias towards female, and so on.

Then in Figure 14b, the statistics for community 2 (profile1) show distinct differences to the general trends. For example, age is predominantly 26-35, and gender has a much greater proportion of female (with respect to whole graph), which corresponds to the profile1 proportion assigned to community 2. Finally, Figure 14c shows the tabular summary of the profile seed assignments vs the real data generated for a community. Again note that the real percentage also depends on the diversity factor, thus the community should have a distinct bias towards the assigned profile1 (when compared with other profiles) but it will not be identical.

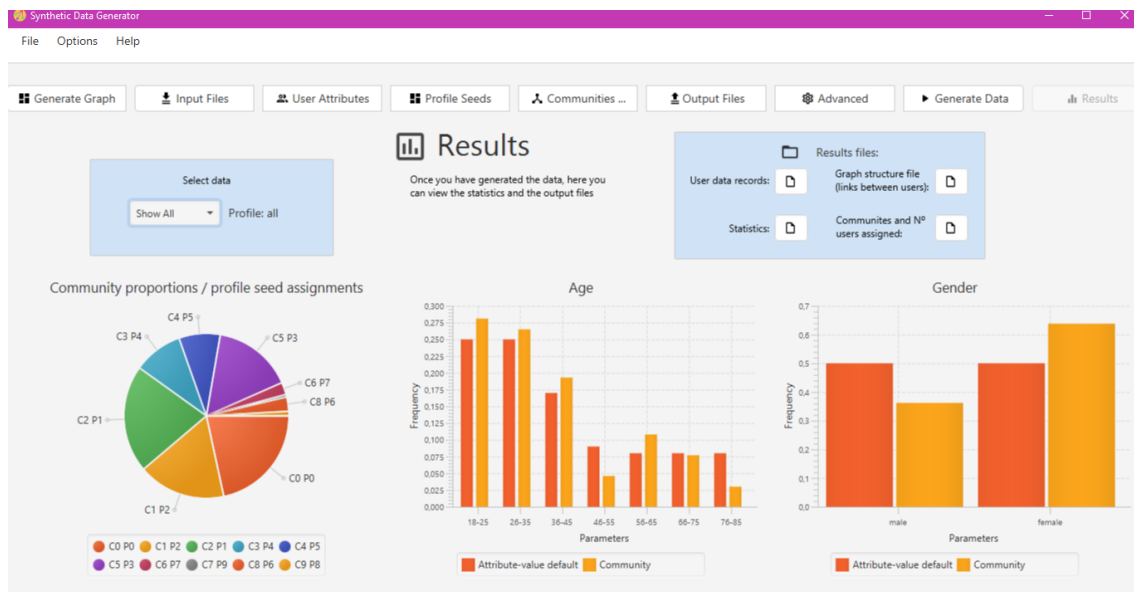


Fig. 14a Statistics Page (all communities)

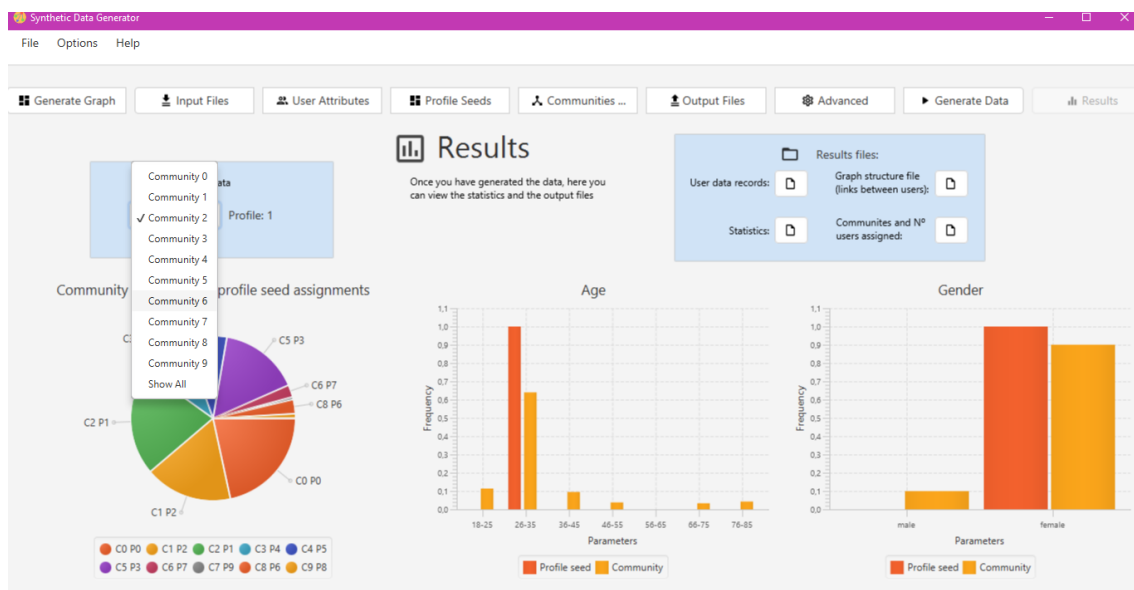


Fig. 14b Statistics Page (community 2)

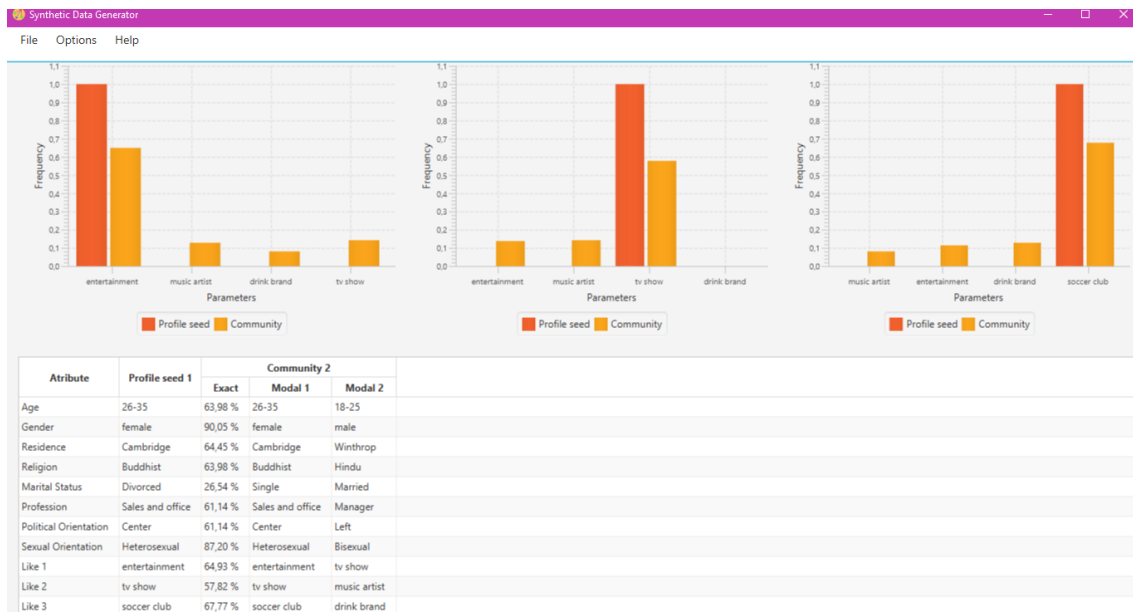


Fig. 14c Statistics Page (community 2)