Viral experiments

Matteo Cristani[(1)], Francesco Olivieri[(2)], Claudio Tomazzoli[(1)]
(1) Department of Computer Science, University of Verona
(2) Data61,CSIRO, Brisbane, Australia
{matteo.cristani, claudio.tomazzoli}
francesco.olivieri@data61.csiro.au

Abstract

When searching for confirming hypothesis on a social domain, scholars regularly form samples to be used for the research. Analogously, for instance, market researchers do. Social networks constitute a huge source of information for these investigations, and it is often used to collect these data. When the collection of the data is made rigorously, an important effort has to be put on the selection of the sample, that should respect criteria of correspondence to the ditribution of the population, balance, and minimality of sample classes, in order to guarantee good inferential statistics evaluation parameters.

In this paper we investigate a systematic way of providing samples on social networks that takes in consdieration the above mentioned criteria *shile forming the sample*. We prove that an affordable and reliable method to construct the basic instrument for the above method (the random table) can de provided, show its computational properties and devise an experimental protocol (the viral experiment protocol) that makes direct usage of the method mentioned above.

1 Introduction

Within social networks a very high number of individuals are active members and participate in the process of generating contents, and collaborating in content generation. At a very general level, we can assume that the population represented by social networks is quite similar to the general population, for the distribution of members can be viewed as pretty similar to the distribution of the population, in a large number of observable variable sets. This assumption is obviously totally false when we consider the world population, and therefore we should assume that social networks per se represent populations of modern industrialized countries, and only partly, developing countries with a high developing rate such as China, India or Brasil. Conversely, large and largely populated southern american and african countries are defiitely not represented correctly, due to the underepresentation of the poorest part of the population that has no access to education and in particular to digital means.

Market researchers are strongly intersted in making their investigations as efficient as possible, in particular by finding ways to generate results of the investigation with

the same accuracy in shortest possible time. They have often make use of on-line questionnaires, that are good means to enshort the process of data gathering for the above mentioned tasks, but require a significant effort of post-processing in the refinement of the sample after the gathering itself, as the data collected are potentially false or meaningless in a number of cases.

It is not difficult to argue that data of the same kind as above are more accurate, and therefore require less post-processing effort, when collected on a social network. In fact, although it is quite possible that a profile on a network is false or inaccurate, this is a rarer occurrence with respect to a simple lie provided on a questionnaire filled in on-line. Publishing a false information, in fact, requires a more delinquent attitude than lying on an anonymous questionnaire. Then, when falsity or omission is detected on a profile, this incoherence can be reported and used to exclude the member from the selection (or the data analysis).

Moreover, the part of data that can be used to settle the class distribution (age, marital status, geographical provenance, language, citizenship) and in some cases, also, to answer some of the questions in the questionnaire (political orientation, sexual orientation, interests) are already present in the profile itself.

Another important aspect is that when forming a sample we need some sort of expression of willingness by the sample subject, and this is often obtained by either retribution, in both economic and other forms, or by social involvement. The second case is often violating the *sample neutrality* hypothesis about the theme. For instance, if we consider a political poll, the political orientation of involved subjects is often *explicit*, whilst when we look at the whole population, the number of persons who actually express their political orientation is very low. Conversely, users of a network can be strongly involved in this processes when responding to *call to action* by other users, in particular by members of the same network in contact with them.

Finally, cross-validation data used to certify quality of a single test, such as repeated questions, or crossed answers among questions, can partly be derived by the profile itself.

The above argument shows that it is perfectly reasonable to consider social networks as good sources of information for on-line questionnaires and this initiatives present the following advantages:

- Part of data are collected from the profile, therefore questionnaires are shorter to fill in;
- Falsities are likely to be less in number;
- Falsities can be detected more easily;
- Involment can be stronger.

Is it possible to perform rapidly the sample formation in a social network? It is quite easy to show that it is more rapid to perform the above mentioned task, when assuming that the involvement determined by social network membership is higher than the direct one obtained in traditional processes of questionnaire subministration.

The rest of the paper is organised as follows. 2 introduces some general definitions regarding social networks and virality, while Section 3 introduces technicalities

regarding viral experiments. In particular Subsection 3.1 describes the proposed process methods for constructing random tables in social netowrks, and Subsection 3.2 discusses the algorithmic solution to the problem of post-selection of the elements of a sample taken from a social network, and finally Section 3.3 discusses the problem of rebalancing the samples, and Section 3.4 illustrates some preliminary experimental results. Section 4 provides references to relevant related work and finally Section 5 takes some conclusions and fiscusses further work.

2 General definitions

A social network is modelled as a graph, with labels on edges and vertices as provided in Defintiion 1.

Definition 1 *A social network is a graph $\langle V, E\rangle$, with labels on vertices $\lambda_V : V \to \Delta_1 \times \Delta_2 \times \cdots \times \Delta_m$, and on edges $\lambda_E : E \to [0,1]$, where labels on vertices represent the user profiles, settled as values in the domain for every observed variable, and labels on edges represent activation thresholds of the start vertex with respect to the target vertex.*

Every domain which we employ to describe the users in the network is equipped with a *null value* intended to represent the fact that the value for that particular member of the network has not been assigned. We assume domains to be finite. Domains are formed by single values, as in for instance, birthplace, that is functional for any member, or by multiple values, as in the case of interests, where an user can have no interests, one single interest, or more than one interest.

The general concept we aim at formulating is that every vertex is described by a profile, and members are connected with an edge that is weighted with the activation probability, namely an abstract measure of the strength of the tie binding the source user to the target user, in the view of the source user.

Example 1 *Consider a social network where the three members are classified based upon their* interests *and* birthplace*. Interest basic values are $\{$sport, politics, cuisine$\}$, and therefore give rise to a combinatorics formed by eight values including the null value, corresponding to no interests.*

Domains are introduced with the explicit purpose of defining the distribution of users as compared to the structured domain obtained by the cartesian product, as in database theory. A structured domain $\Delta = \Delta_1 \times \Delta_2 \times \ldots \times \Delta_m$ represent the population on which the concept of *population distribution* and the one of *sample* are defined. Given a domain Δ of dimension m, a set $\mathscr{C}$ of relevant classes on Δ is formed by n subclasses of Δ. A *distribution of probabilities* on a set of n relevant classes is a probability vector of dimension n, where, for each class C_i, the probability p_i is the ratio $\frac{\|C_i\|}{\|\Delta\|}$.

We can define a sample on a domain only when we are able t define the distribution on the population. Therefore we assume that a domain Δ is provided as associated to a

distribution on the population itself, for a given set of classes of interest of the investigation. A class on a population is a logical condition on the domain itself, typically a query on the structured domain. Abstracting away from the form of the query language, we can assume that the classes are sets of values in the structured domain, namely a class C is $C \subseteq \Delta$, and for every class C we name q_C the query on Δ that generates C, and by $q_C(X)$ the intersection between C and X.

We thus define the samples as in definition below.

Definition 2 *Given a structured domain Δ, with m classes $C_1, C-2, \ldots, C_m$, a subset P of Δ, and a threshold τ, P is a* sample *on Δ with accuracy of distribution τ, when, for every class $C \in P$, generated by the query q_C, the following holds:*

$$|\frac{\|C_i\|}{\||\Delta\|} - \frac{\|q_C(P)\|}{\|P\|}| \leq \tau$$

The Bayes rule on the determined distributions holds on each of the defined conditions C as the opposite condition $\neg C$ has the probabilistic complement value on the domain. The τ threshold of accuracy can be easily show to hold on the complement class as well.

3 Viral techniques for social experiments

The construction of a sample occurs in two steps, in every sociological experimentasl setting. Among the population, a large number of elements is selected that respond to two basic requisites: they are numerous enough, by a predefined minimum size σ for each class of the sample, and they respect the distibution of the classes by a predefined tolerance τ as mentioned above.

In the definition phase, the selection of the members in the population usually follows a golden rule of *involvement* by either paying the sample subjects or because they participate in the experimental setting itself. The golden rule above has a few major drawbacks:

- Sample members should not be aware of what the experiment aims at proving more than they should when the test is performed in a medical setting or pschological ones. In fact, the knowledge of the purpose of the experiment by the subjects can produce the tendency of these to respond to the test in an oriented way, making the experimental setting not neutral;
- Paid subjects respond more easily to the involvement when they are in need of the money used to hire them for the experiment, and therefore the richest part of the population is generally not mapped as appropriately as the poorest one, when this approach is used;
- Budget shortage, that can be very common in scientific communities, and is a limit to market investigations for small companies, can be the reason for quite small samples, or, more formally, to the construction of samples with a minimum size parameter very low.

- Involving people in environments where they are interested in the topic can produce a shift in the orientation. For instance, an investigation that aims at establishing the quality of vegan food cannot be spread among the population that is required to participate to the selection process expressing their interest in the tests. Either the test involves eating meat or fish, and thus vegan people will not participate, or it is involving vegan people in the construction of the sample itself, making the sample oriented in favor of vegan food quality.

The above drawbacks can be partly overwhelmed if we retrieve the subjects on social networks, and involve them, for free, in the experiment itself, without specifying the scientific hypothesis that is to be verified, but only the area of the investigation. Once we have provided such pre-settled aspects, a sociological investigation can be perfomed on a sample obtained by social networks.

To define correctly the above mentioned framework we need to have the following hypotheses:

- Members of the network have to be marked by labels representing the values to be attributed to the variables used to classify the members themselves;
- Call-to-action are messages passed by one member of the network to another one and the willingness to answer to such a message is determined by the social relationship between the sending member and the receiving member;
- Once a member has been involved the messages representing calls-to-action are spread out even if the member decided not to participate in the sample.

To represent the above mentioned constraints, and define correctly the framework within which we settled the notion of viral experiment, we need to formulate three notions. First of all, we aim at describing a $\mathscr{SN}$ at a very abstract level, where the members of the network are mapped directly onto *vertices* of a graph, while an edge connecting one vertex v_1 to another vertex v_2 represents the abstract relationship that v_1 *can communicate to* v_2.

Definition 3 (Social network) *A* basic social network *($\mathscr{BSN}$) is a graph $S = \langle V, E \rangle$, where V is a finite set of* members *(the vertices) and E is a relation on V (the edges), whose elements are* communication channels*. An edge connects the source vertex to the target vertex and the intended meaning is that the target vertex receives messages from the source vertex by the communication channel established by the edge.*

For instance, when on Twitter or Instagram an user y *follows* an user x, we represent this by an edge between x and and y, giving account to the fact that when x publishes something, then y receives some communication about it.

More specifically, we consider a labelling of the above defined graph with one label λ_r^v on each vertex and three distinct labels λ_a^e, λ_{min}^e and λ_{max}^e on each edge, where:

- λ_r^v is the *reception time* for a vertex v, which is either *null* (when the vertex has not yet been reached by the message) or a positive integer, and represents the instant when the message arrives to the member. The label 0 is reserved to

represent the situation in which a vertex is a *seed*, namely when the message has been delivered to the vertex from outside the network itself, and the message delivery starts, actually, from there;

- λ_a^e is the *activation threshold* on edge e, a real number such that $0 \leq \lambda_a^e \leq 1$, used to denote the probability that the member sends the message she has received on communication channel e, at every admissible instant of time;
- λ_{min}^e is the *minimum latency time* on edge e, a positive integer representing the time span passed before the message can be send out;
- λ_{max}^e is the *maximum latency time* on edge e, a positive integer representing the time span after which the message cannot be sent out anymore.

Vertex v will tries to send out the message on edge e in between the interval $[\lambda_r^v + \lambda_{min}^e, \lambda_r^v + \lambda_{max}^e]$. Clearly, not every labelling of the vertices is coherent with the labels on the edges. Determining this coherence is a trivial task, therefore, without loss of generality, we assume here that we only manage coherent labelings.

We make some obvious assumptions, namely that the minimum time span has to be lower than the maximum time span, that the probability label cannot be zero (the zero probability on an edge is represented by the absence of the edge).

For an edge (v, v') we call v the *source vertex* and v' the *target vertex*.

The above labeling, that we name *behavioral* does not represent the labeling of the aspects of the users that are employed to provide the mapping. We name this labeling *profile*.

A profile is a label on vertices that associates each vertex to a vector $\sqsubseteq$ on the structured domain $\Delta = \Delta_1 \times \Delta_2 \times \cdots \Delta_m$ and represents the values assumed to describe the members.

A *seeding* P of a $\mathscr{SN}$ is a partial labelling of the vertices with initial temporal instants, namely 0. A seeding consists in the selection of a set of agents aimed at delivering the message first.

Seeds, as well as members involved afterwards in the process of distributing the message on the network, are associated to a probability of involvement in the content of the message distributed, namely to the call-to-action that is an external trigger to the members themselves. We can look at this process as defined in Example 2

Example 2 *A sociologist aims at proving the following hypothesis: young italians are more interested in politics than elder ones. To prove this hypothesis she provides an online questionnaire on a website and launches a campaign that is stated to investigate the interest of the italian population in politics, without revealing the specific orientation of the hypothesis. The campaign is launched on a social network, for instance Facebook, and it consists in sending messages to a group of seeds who can be willingful to participate to the passive phase (the spread out of the call-to-action) and to the active phase (the filling of the online questionnaire) with significantly different probabilities. Passive involvement probability is 30%, whilst active involvement probaility is 10%.*

To evaluate the needs for the above settled experiment, our fictious sociologist need to forecast:

- Time for completing the spread-out on the entire network, as a limit to the campaign itself;
- Time for forming a sample with correct distribution with respect to the relevant classes, and relatively to the minimum size of each class.

We can thus provide some basic problems in the above defined framework, that are interesting for the construction of a correct solution by a sociologist.

Definition 4 *Given a social network $\mathscr{SNS}$, labelled on the vertices with a profile labelling coherent with classes $\mathscr{C} = \{C_1, C_2, \ldots, C_m\}$, a minimum size of classes μ, a probability of active involvement π, a probaility threshold ξ, and finally a seed σ on S, we name* sample selection direct problem *the problem of establishing whether it is possible to reach all the vertices of S with a probability greater than or equal to ξ, while starting from σ, respecting the minimum size constraint for each of the classes in $\mathscr{C}$ thanks to the active involvement of enough members of the network. The* sample selection inverse problem *can be defined as above with the inverse request of determining one seed, if one exist, that satisfies the mentioned requests.*

The probablity of passive involvement is determined by the usage of classical Bayes *a posteriori* probability computation. Every vertex x is entered by edges labeled by a probability that represents the willingness of transmission by the source vertex. The probability of passive involvment is determined as the probabilistic complement of the product of probabilistic complements of those probabilities, at every instant of time. Now, if we assume that the transmission of the message has the same probability in every instant of time between the minimum and the maximum latency time, this probability evolves by following the same model of probabilistic complement along time. In other terms, the proability that the vertex is reached at time t can be viewed as the sum of the complement of the probability of being reached at time $t-1$ and the above mentioned Bayes formula. The complete formula appears in Equation 1

$$p(v,t) = p(v,(t-1)) + (1 - p(v,(t-1))) \cdot \prod_{e \in E(v,t)} [1 - (p(s(e)) \cdot p(e))] \qquad (1)$$

In Equation 1, we refer by $p(v,t)$ the probability that the vertex v is reached by the message not later than time instant t, by $E(v,t)$ the set of edges that connect the vertices $v' \in E(v,t)$ to the vertex v, that are *active* at time instant t, namely such that the time passed since the message reached firstly v', at temporal instant t, is included between the minimum latency time and the maximum latency time, on the edge connecting v' to v. The message has been transmitted from the seeds to the other members of the network and, step by step, has reached a subset of the vertices that are connected to v. By $p(s(e))$ we denote the probability label on the source edge of the edge e and by $p(e)$ we obviously denote the activation threshold on the edge e.

Once we have computed the passive probabilities as defined above, the second step of the method consists in computing directly the product of passive probabilities with the active probabilities on each vertex, and finally, in computing the product of the probabilities obtained as defines above on all vertices. The obtained proability is the probability that the coverage has produced enough elements of the sample, and therefore is a solution of the proposed direct sample selection direct problem. On this basis we can easily prove Theorem 1.

Theorem 1 *The sample selection direct problem is polinomially solvable in $O(n^3)$ with n number of vertices of the network.*

Thanks to Theorem 1 we can instead formulate Theorem 2.

Theorem 2 *The sample selection inverse problem is polinomially solvable in NP-hard.*

The proof is not reported for the sake of space, and is a straightforward polinomial reduction of a simplified form of the mentioned problem to the inverse coverage of directed graphs. Clearly this does not prove the NP-completeness, as we need to provide a polinomial solution on deterministic machines. To do so, we need before to prove polinomiality of the direct problem. Then we shall prove that it is possible to provide a non-deterministic algorithm that solves the problem by invoking an oracle on the above polinomial solution.

By means of the technique sketched above we are able to provide a forecast of the success of a campaign. We can use the method to estimate the probability of reaching the entire network (and clearly, by fixing a single threshold on the vertices, we can invert the method and compute the probability that the number of actively involved members is greater than a given probability). The methods can be used preliminarly to a sociological experiment in order to provide room for programming the experiment itself. The second phase of a viral experiment campaign is devised in Section 3.1.

3.1 Random tables in social networks

Once the campaign has been launched, the people involved in the process fill in the questionnaire and provide data for the sociologists. As shown in [11] and [9], the profiles can contain interests, and interests drive homophilic connections, that favor message flows. This aspect suggests one basic seed selection criteria, that can be used as a means to decrease in a significant way the complexity of the average case for the above mentioned inverse sample selection problem. In this section we shall formulate a few criteria of the same type, that are used to build up an heuristics useful for the mentioned issue.

Criterion 1 *The seeds shall provide a coverage of the variance of values in each domain.*

It is also very convenient that, in the hypothesis of relatively dense networks, for instance with mean distance between vertices of three, the seeds are quite set apart, in terms of shared connections, in order to guarantee a rapid convergence to the coverage, if any. This can be done by again using homophilic distance measures.

Criterion 2 *The seeds shall be as distant as possible to each other.*

First experiments show that the usage of the above mentioned criteria alone is not enough to reduce significantly the cost of the selection of seeds. It is however rather natural to consider that seeds satisfy the following two criteria.

Criterion 3 *The seeds shall be as likely to transmit messages.*

Criterion 4 *The seeds shall be as likely fill in questionnaires.*

Again, first experiments show better performances with the four criteria, but still do not bring to a significant improvement in the average case. The second criterion, however, suggests a method for seed selection that is positively solving the issue of improvement, again in the experiments we settled (very preliminary) that we illustrate below. The idea behind this criterion is to select individuals who are as close as possible to other individuals who have high probability of being activated in filling in the questionnaire. This is a simple extension of the *closeness centrality* but limited to the computation on vertices with high probability of active involvement.

Criterion 5 *The seeds shall have high closeness to vertices with high probability of active involvement.*

A very natural method to provide good seed sets is random extension. We simply select a reduced seed set, and extend the set with new elements satisfying the above mentioned criteria.

If we combine the mentioned criteria with the heuristic method illustrated above, we obtain good performances, in the experimental settings we performed the tests on. We have involved four experimenters with their egonets (the networks formed by their contacts) in FaceBook and WhatssApp, to test the sociological hypothesis that the probability of being involved in an argument with somebody else is proportional to the homophily in the topic of the argument. A part the result of the investigation that is not central to this discussion, we have some 1321 profile descriptions, and more than 10.000 messages, posts, comments, and threads. The application of the criteria to these ex-post data have shown that the hypothesis of reduction of the computational cost of seed selection is realistic.

Once we have formed a sample, the sample itself requires to be rebalanced. To do so, we need first to rank elements in the sample classes.

3.2 Ranking samples

Ranking criteria of members of social networks have been developed widely in the recent past, starting with the development of specific new social network analysis methods evolved from classic ones, such as degree centrality, closeness centrality and betweenness centrality, possibly mitigated or modified by the introduction of semantic-based measures.

The idea of the selection process here is to try to detect those members of the network who, while satisfying the selection criteria for their class, are also likely to satisfy further important aspects:

- We should prefer members who exhibit values of the measures in the distribution *closer* to the median. These members are less likely to perturbate the final result of the research in terms of class distribution.
- We should prefer members who are unlikely to be fake. Evidently this criterion has to be applied before that members are included in the class, to prevent them to shift the results of the research. However, once we have established a threshold to exclude members who are likely to be fake, this further criterior applies to the ranking as well.

Clearly, ranking operations are more effective when applied to larger classes, and therefore it is important to regulate post-selection phase on the size of the sample classes. A possible method to do so is the re-balancing approach. Central to this method is the cycle that is obtained by it, that is the final outcome of our investigation: a methodology for the selection of members of a network for sociological experiments.

3.3 Rebalancing samples

When we rank elements of the classes that have been selected we have a final result formed by elements of the classes in the sample. Obviously, these are not balanced, and a simple process for rebalancing it consists in ranking the elements, and choose as much elements as required by the criterion of minimal amount of elements for each class. However, quality of this process is questionable, when distribution indices such as standard deviation or variance are quite high. In fact, when size of classes differ significantly, it might be the case to correct the above sketched selection process, consisting in selection, ranking and post-selection. In particular there is one rather simple case in which the correction is needed: when the ranking operation gives out a completely different statistical behavior in classes with fewer members than in classes with more members. In this case, it is worth adding elements to the smaller classes until an equilibrium in the distribution of ranks in the classes arises.

To do so, we propose an approach, consisting in iterating the selection process until we reach a number of members that satisfies the criterion of having variance of relative rank distances between minimum and maximum values in the classes that is under a given threshold. This is satisfied by continuing the search for members able to integrate the classes under the same selection criteria defined in the sample selection problem and adding this rebalancing value as a parameter.

The general schema can be defined as in Algorithm 1. We assume that we already implemented an algorithm for computing the coverage, namely able to iterate in a simulation the process of delivering the call-to-action so that the message reaches those who are possibly involved in the filling in of the questionnaire. We iterate this until the sample calsses are filled. After that we have a second step consisting in ranking. We then go on with an *anytime* fashion with the first two steps until the number of members in the classes satisfy the parameter of variance of the input. The number of members of each class is then settled to the minimum satisfying the ranking requests, that miht be higher the minimum for the general problem without rebalancing.

The rebalancing process ends up when the result is obtained and it is perfect. It is easy, however, to show that at each iteration, the algorithm produces, in the worst

Data: A $\mathscr{SN}$ S, a probability of active involvement threshold τ, a variance threshold v, a minimum number of members of classes $\ni$, a vector of m classes C, with associated probability for each class $p(C_i)$, and for each class a logical condition $\chi(C_i)$, establishing belonging to C_i, and a ranking function r

Result: C', a set of classes satisfying the same criteria used for C, each formed by at least n members, where the distribution over the classes in C' is identical to the distribution on C within a percentage of discrepance less than τ for each class, and satifying also that the distribution of the percentages of the classes has a variance of rank values within v.

repeat

  Distribute message on the network;

  Add filled-in questionnaires to the sample database, only for elements of the classes that were not balanced at the step before, if any;

  Rank elements in the sample classes.

until *Rank rebalancing completed*;

Algorithm 1: Sample selection process.

case, a set of elements that is balanced exactly as it was in the step before. Therefore, stopping the algorithm in any instant of time gives you an approximation that is worse or equivalent to the one you obtain if you interupt it afterwards. This characteristic allow us to prove Theorem 3.

Theorem 3 *Algorithm 1 is an anytime method.*

The complexity of the above method is an easy consequence of the definition of the problem.

Theorem 4 *Algorithm 1 performs in $O(n^3)$, with n number of vertices in the network.*

3.4 Preliminary experimental results

We conducted a very simple experiment, based on the usage of the egonet of four human experimenters. Table 1 illustrates the results. The experiment consisted in driving members of their egonets on FaceBook of the four mentioned subjects onto the request of filling in a simple satisfaction questionnaire about their school activities. The questionnaire simply consisted in three questions:

- Did you have any activity proposed by the school, extra curriculum?
- Did you attend some of these activities?
- Do you consider youself satisfied with those activities?

The samples have been formed to respect the distribution Male/Female (with the classes considered perfectly balanced) and a tolerance of 5%. The accepted variance

Egonet	Number of elements in the egonet	Selected items	
Student1	131	19	25
Student2	214	17	24
Student3	184	22	28
Student4	204	19	22

Table 1: Preliminary results in the selection of samples by Algorithm 1.

Egonet	Male	Female
Q1	37%	39%
Q2	17%	23%
Q3	94%	89%

Table 2: Questionnaire results.

was 2%, the ranking method was the degree centrality, and the subset minima were 15. The result have been rebalanced by the algorithm in a complete fashion. Results of the test on the questionnaire are on Table 2. As the reader can easily notice, the number of involved subjects of the egonets is limited, but the results are quite promising as the rebalancing method has given a very nice sample selection result. In Table 1 the selected items correspond to responding subjects, namely those subjects that have been involved in the process, and selected for the questionnaire, and then ranked and accepted after rebalancing. The subjects on fourth column are instead those who have been involved on first step, namely responding subjects who have not been post-selected either for eliminating redundancy by ranking or by the rebalancing process. In table 2 we report number of responding subject who give positive answer to the three above questions.

4 Related work

Numerous investigations have been carried out in several different communities regarding the ways in which messages are delivered through networks, and some recent studies have also found ways to combine these efforts [17]. In the investigations about transmission protocols (see [14] for the comparison of the most common network models), the notion of *transmission delay* has been investigated in many different and diverse contexts [23]. On the one hand, the above mentioned literature is rather important to this investigation, as it has provided the reference to notion of transmission probability, and has been the source of the preliminary definitions provided in this paper. On the other hand, it is much more important to study the results obtained in the Social Network communities, in particular investigations on the theoretical aspects of virality in networks, applications to social networks [3, 22, 15], and many works on opportunistic evolution techniques [16, 1, 13].

There is also a thriving literature about experiments on social networks, but not many investigations have focused upon the specific, but central, problem of sample se-

lection. In a recent research [19] Parker et al. have studied the relationship between social network and social connection in the design of experiments. We can refer this and other two as basic references of our investigation. A second important reference regards the notion of active participation in social networking, in particular for commercial activities [5].

A case study in medical research have been proposed by Pourabbasi et al. [20] that invetigates diabetes. On a completely different viewpoint we find a research applied to economics [4] that has been the logical basis for our distinction between passive and active threshold measures.

On a general side we can find studies that prove ability of social networks to influence positively the result of investigations in terms of quality [21].

We follow-up several studies about notions of collaborativity that have been defining attitude to cooperate in social networks as mirrors of the social attitudes in general from many different viewpoints [6, 7, 10, 2, 12, 18]. We firstly initiate here an investigation with experimental basis.

5 Conclusions

In this paper we dealt with the problem of defining a general framework for performing research market and sociological scientific investigations on social networks, supported by formal results regarding the computational costs of seed selection, sample formation, sample rebalancing.

There are at least three fundamental issues that deserve to be considered further. First of all, we need to provide a security belt on the selection of seeds and in general in the acceptance of members of networks as reliable in answering to an on-line questionnaire. The basis for doing this is to recognize the situations in which it is possible that profiles are fake, or partly dissimulated, as, for instance, in [8], and to detect intrinsic falsness of questionnaire answers as often done by sociologists by using specific methods (repeated questions). Finally, it is important to prevent proper attacks, like fake identity dissimulation, man in the middle in the online questionnaire systems, or even worse, bot questionnaire fillers.

Secondly, we need to provide room for methods to collect profile data from *more than one* social network, to give the sociologists the opportunity of mapping people in a more accurate and complete way.

Finally, we are attempting at studying methods for *follow-up* questionnaires, for instance, market researches on satisfied customers, usually driven on some sort of retribution, and therefore demanding control of effectiveness and efficiency.

References

[1] M.W. Bigrigg, K.M. Carley, K. Manousakis, and A. McAuley. Routing through an integrated communication and social network. In *MILCOM - Proceedings*, 2009.

[2] E. Burato and M. Cristani. The process of reaching agreement in meaning negotiation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7270 LNCS:1–42, 2012.

[3] P.-Y. Chen, S.-M. Cheng, and K.-G. Chen. Optimal control of epidemic information dissemination over networks. *IEEE Transactions on Cybernetics*, 44(12):2316–2328, 2014.

[4] L. Corazzini, F. Pavesi, B. Petrovich, and L. Stanca. Influential listeners: An experiment on persuasion bias in social networks. *European Economic Review*, 56(6):1276–1288, 2012.

[5] R. Cordero-Gutirrez and L. Santos-Requejo. Intention to participate in online commercial experiments by social networks users: Differences in gender and age. *Management Research Review*, 39(4):378–398, 2016.

[6] M. Cristani and E. Burato. Modelling social attitudes of agents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4496 LNAI:63–72, 2007.

[7] M. Cristani and E. Burato. A complete classification of ethical attitudes in multiple agent systems. volume 2, pages 1122–1123, 2009.

[8] M. Cristani, E. Burato, K. Santacá, and C. Tomazzoli. The spider-man behavior protocol: Exploring both public and dark social networks for fake identity detection in terrorism informatics. volume 1489, pages 77–88, 2015.

[9] M. Cristani, D. Fogoroasi, and C. Tomazzoli. Measuring homophily. volume 1748, 2016.

[10] M. Cristani, E. Karafili, and L. Vigan. Blocking underhand attacks by hidden coalitions. volume 2, pages 311–320, 2011.

[11] M. Cristani, C. Tomazzoli, and F. Olivieri. Semantic social network analysis foresees message flows. volume 1, pages 296–303, 2016.

[12] G. Governatori, F. Olivieri, E. Calardo, A. Rotolo, and M. Cristani. Sequence semantics for normative agents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9862:230–246, 2016.

[13] S.-L. Huang-Fu, B. Guo, Z.-W. Yu, and D.-S. Li. Incentive mechanism for opportunistic social networks: the market model with intermediaries. *Ruan Jian Xue Bao/Journal of Software*, 25:53–62, 2014.

[14] H. Inaltekin, M. Chiang, and H.V. Poor. Average message delivery time for small-world networks in the continuum limit. *IEEE Transactions on Information Theory*, 56(9):4447–4470, 2010.

[15] K. Jahanbakhsh, V. King, and G.C. Shoja. Predicting missing contacts in mobile social networks. *Pervasive and Mobile Computing*, 8(5), 2012.

[16] S.K.A. Khan, R.J. Mondragon, and L.N. Tokarchuk. Lobby influence: Opportunistic forwarding algorithm based on human social relationship patterns. In *PERCOM Workshops*, pages 211–216, 2012.

[17] Z. Lu, Y. Sagduyu, and Y. Shi. Friendships in the air: Integrating social links into wireless network modeling, routing, and analysis. In *INFOCOM - Proceedings*, volume 2016-September, pages 322–327, 2016.

[18] F. Olivieri, G. Governatori, S. Scannapieco, and M. Cristani. Compliant business process design by declarative specifications. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8291 LNAI:213–228, 2013.

[19] B.M. Parker, S.G. Gilmour, and J. Schormans. Optimal design of experiments on connected units with application to social networks. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 66(3):455–480, 2017.

[20] A. Pourabbasi, J. Farzami, M.-S.E. Shirvani, A.H. Shams, and B. Larijani. Using virtual social networks for case finding in clinical studies: An experiment from adolescence, brain, cognition, and diabetes study. *International Journal of Preventive Medicine*, 8, 2017.

[21] D.G. Rand, S. Arbesman, and N.A. Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences of the United States of America*, 108(48):19193–19198, 2011.

[22] H. Sun and C. Wu. Epidemic forwarding in mobile social networks. In *ICC - Proceedings*, pages 1421–1425, 2012.

[23] Y. Zhu, H. Zhang, and Q. Ji. How much delay has to be tolerated in a mobile social network? *International Journal of Distributed Sensor Networks*, 2013, 2013.