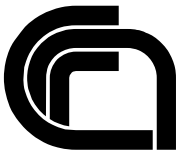

*Consiglio Nazionale delle Ricerche*

# Generating Social Network Graphs

M. Conti, A. Passarella, F. Pezzoni

IIT TR-22/2011

**Technical report**

**settembre 2011**

**Istituto di Informatica e Telematica**

# Generating Social Network Graphs

Marco Conti, Andrea Passarella, Fabio Pezzoni CNR-IIT, Via G. Moruzzi, 1 - 56124 Pisa, Italy
{m.conti, a.passarella, f.pezzoni}@iit.cnr.it

**Abstract**

In this paper we present and evaluate a social network model which exploits fundamental results coming from the social anthropology literature. Specifically, our model focuses on ego networks, i.e., the set of active social relationships for a given individual. The model is based on a function that correlates the level of emotional closeness of a social relationship to the time invested in it. The size of the social network is limited by the time budget a person invests in socializing. We exploit the model to define a constructive algorithm to generate synthetic social networks. Experimental results show that our model satisfies, on average, known properties of ego networks such as the size, the composition and the hierarchical structure. We also introduced a procedure for the integration of different ego networks and the generation extended social networks.

**Index Terms**

social networks; ego networks; model;

## I. Introduction

The emerging pervasive and social networks are drastically changing the (information) society. First of all, we are experiencing a convergence between the cyber/virtual and physical worlds. The convergent cyber/physical world will be *content-centric* where content generated in the physical space is immediately transferred to the cyber space (e.g., multimodal sensing), and cyber outcomes have immediate impact on physical space. Humans are at the core of this convergence; each person has several (mobile) devices through which he/she can interact with the virtual world thus linking the physical world and the electronic world of users devices [1]. In this scenario, *human and online social networks* have a very important role for accessing and circulating the massive scale of content that is circulating in the network/society. By translating human relationships in the electronic world, we embed in electronic devices the key characteristics that enable humans to effectively handling and sharing large amount of information.

Human relationships can be exploited in the virtual world for fast and effective circulation of data with spatial temporal value and for content provision and personalized context, such as by sharing information of mutual relevance.

There is significant evidence suggesting that *human* social networks (i.e. the set of social relationships people maintain with each other) are not particularly affected by specific communication technologies [2]. Therefore, it is reasonable to see the properties and structures of human social networks as an invariant with respect to the evolution of the underlying means supporting social interactions.

Human social networks exhibit remarkable dynamism and structural properties that may significantly affect the quality of the information (i.e., trust and reputation, relevance, reliability, etc.) and the way information may circulate; it is conjectured necessary to transverse only a small number of human social relationships in order to connect *any* pair of people resulting in the "small world concept". Therefore, understanding and modeling human social networks is a fundamental step in designing efficient protocols for data dissemination in the cyber-physical world. In this paper we present a first important step in this direction. Specifically by exploiting social anthropology results we have developed a model of the ego network, i.e., the model describing the set of active social relationships of an individual. Results from Dunbar et al. [3], [4] indicate that human relationships have a hierarchical structure and, on average, an individual has up to 150 active social relationships, i.e., the Dunbar number. These results constitute the bases for the model developed in this paper.

The properties of the ego networks are summarized in the next section and our model is presented in Section III. In Section IV we define the functions and the parameters that characterize the model. In Section V we validate the model and formulate the conclusions while in Section VI we introduce a procedure for the integration of different ego networks and the generation of extended social networks.

## II. Ego Networks

Ego networks are a particular category of social networks made up of an individual (an "*ego*") and the people ("*alters*") with which the ego has some kind of social relationship.

There are limits to the amount of social relationships that an individual can maintain, this is due to cognitive and time constraints [5]. In fact, keeping social relationships demands cognitive resources and time available to invest on them and both resources are limited. Different studies about ego network size have been conducted (e.g. in [4], [6], and [7]). It has been demonstrated that ego networks have a hierarchical structure that consists of a series of concentric layers of acquaintanceship with increasing sizes. Dunbar et al. suggests that the layers in an ego network are: "*support clique*", "*sympathy group*", "*band*" and "*active network*" (the whole network) with sizes $\sim 5$, $\sim 12$, $\sim 35$ and $\sim 150$ respectively, [3], [4]. The layers are hierarchically inclusive, so that each layer includes all inner levels. This structure is depicted in Figure 1. Sometimes in this paper, we use the term *external part* of a layer in order to refer to the part of the layer not overlapped with its inner levels.

Support clique and sympathy group are made up by a relatively small number of alters the ego is emotionally closest to. On the other hand the alters connected to the ego by weak ties, which represent the greatest part of the network, are included in the external layers. Each layer of the network has specific characteristics: support clique and sympathy group are well-defined in size and composition (see [8] and [9]) as well as the active network is ([4]), while no accurate information is currently available in literature about the band level. Therefore in this paper we do not explicitly model the band level and we consider it merged within the active network layer.

Regarding the correlations among the layers' sizes, the study in [8] suggests that there is a linear correlation between support clique and sympathy group. On the contrary there is no information in literature about possible correlations of their sizes and the size of the active network layer.

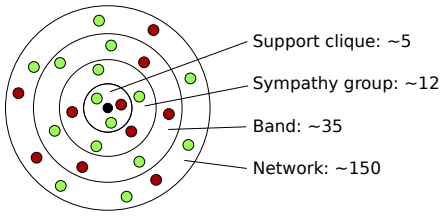


Fig. 1. Hierarchical ego networks' structure. The black circle represents the ego; dark red circles refer to the kin; light green ones refer to non-kin.

Relationships in social networks may be classified into different categories such as: kin, friends, neighbors, work colleagues, etc.. Moreover alters may be characterized by their gender, age, education level, marital status and so on. In social networks each relationship is also characterized by a level of *emotional closeness*. Strong relationships have a higher level of emotional closeness compared with weak ties. As suggested by Hill & Dunbar, the emotional closeness level may be the key parameter to consider in order to select in which layer a relationship has to be included [4].

The level of emotional closeness is positively correlated with the "frequency of contact", which is estimated with the inverse of the "time since last contact" [4]. The latter also reflects the time invested in a particular relationship [5], therefore it is generally assumed that there is a relation between the time invested in a relationship and the level of emotional closeness. Maintaining a relationship at high level of emotional closeness requires a lot of time invested in it, for both friends and kin. On the contrary, for low levels of emotional closeness, kin relationships require less invested time than the relationships with friends [5].

## III. The Model

Our model allows to define ego network graphs that, on average, have the properties described in literature, such as the size, the hierarchical structure and the composition of each layer. The model is based on parameters and functions, defined in Section IV, that are obtained exploiting results in the reference literature about the average ego network.

As previously said, the size of the network is limited by cognitive and time constraints. Since cognitive constraints are not easily quantifiable, our model focuses on time constraints, associating to each ego a certain time budget for handling his/her social relationships. In the model each relationship requires a specific amount of time, therefore the size of the ego network is constrained by the time budget.

In order to know the time requested by each relationship, the model exploits a function that, given the level of emotional closeness of a relationship, returns the related amount of time to handle it. The level of emotional closeness is distributed according to known probability distributions, and identifies the layer a relationship belongs to. Each layer of the ego network is related to specific interval of values of emotional closeness. The function that correlates emotional closeness and time is defined in order to obtain, on average, networks with a specific expected size.

The size of the sympathy group follows a specific distribution and it is independent of the network size. On the contrary it is linearly correlated to the support group according to a ratio defined by a probability distribution.

As previously said, the literature proposes different categorizations of relationships and alters. Our model only considers the kinship with the ego and the gender of the alters because there are many data available about these categories [8], [4]. Therefore, each relationship in the model is characterized by the type (kin or not-kin) and by the gender of the alter according to the composition of an average ego network. Our model simply considers static ego networks. Including the evolution over time, studied in [10], represents an interesting future work.

In the following subsection we present an algorithm for the generation of ego network graphs that are based on the presented model.

### *A. The Algorithm*

The algorithm generates an ego network graph iteratively, following the proposed model. It adds relationships to the network from the inner to the outers layers, until the time budget is completely spent. To construct the ego graph, the algorithm exploits a set of functions ($h_d$, $f_S$, $f_W$, $f_B$, $f_{A,D}$ and $f_E$) and parameters ($\mu_l$ and $m$) whose values are obtained in Section IV from the analysis of an average net.

The first step is the creation of an ego and the assignment of its gender. The gender of the ego is saved in the variable $g$ that can take values $\mathrm{M}$ (male) and $\mathrm{F}$ (female). The algorithm extracts $g$ form a Bernoulli distribution $Ber(m)$ where $m$ is the probability that $gen = \mathrm{M}$ (line 2-3).

The next step is the extraction of the sympathy group size $s_{\mathrm{sym}}$ from the known probability density function $f_S$ (line 4). The mean value of $f_S$ is $\mu_{\mathrm{sym}}$ that is the size we expect to obtain, on average, by the algorithm execution.

Once the algorithm knows the value of $s_{\mathrm{sym}}$, it can obtain the size of the support clique $s_{\mathrm{sup}}$. In order to do this the algorithm randomly extracts the ratio $w$ between the two layers' sizes from the density function $f_W$. Once $w$ is extracted, the algorithm sets $s_{\mathrm{sup}} = s_{\mathrm{sym}} \cdot w$ (lines 5-6). The expected value of $s_{\mathrm{sup}}$ is $\mu_{\mathrm{sup}}$.

Since the probability density functions used in the model return continuous values, but layer sizes have to be natural numbers, values are rounded using the dithering method [11]. Moreover each negative value is converted into a zero.

In the next step the algorithm assigns the time budget $bdg$. This amount is extracted from the known probability density function $f_B$ (line 7).

At this point the main loop starts (lines 9-22). For each iteration the algorithm sets the parameters for a new relationship that is created only if there is enough time available. The total amount of time spent by the created relationships, is kept in the variable $tot$, that is updated after each relationship addition. The variable $tot$ is initialized before the loop begins together with the control variable $done$ and the counter $i$, which represents the current size of the network (line 8).

By the knowledge of the current size $i$ and the layer sizes $s_{\mathrm{sup}}$ and $s_{\mathrm{sym}}$, the algorithm infers the current layer $l$. The variable $l$ takes the values in the set $\mathcal{L}$: $\mathrm{sup}$ (support clique), $\mathrm{sym}$ (sympathy group) or $\mathrm{net}$ (active network) (line 10).

**Algorithm 1** Ego Network Creation

1: **procedure** CREATEEGONETWORK
2: $g \leftarrow$ EXTRACTFROM$(Ber\,(m))$
3: $ego \leftarrow$ CREATEEGO$(g)$
4: $s_{\text{sym}} \leftarrow$ EXTRACTFROM$(f_S)$
5: $w \leftarrow$ EXTRACTFROM$(f_W)$
6: $s_{\text{sup}} \leftarrow s_{\text{sym}} \cdot w$
7: $bdg \leftarrow$ EXTRACTFROM$(f_B)$
8: $done =\leftarrow$ `False`, $tot \leftarrow 0$, $i \leftarrow 0$
9: **repeat**
10: $l \leftarrow$ SELECTLAYER$(i, s_{\text{sup}}, s_{\text{sym}})$
11: $a, d \leftarrow$ EXTRACTFROM$(f_{A,D|L=l,G=g})$
12: $e \leftarrow$ EXTRACTFROM$(f_{E|D=d}\ in\ (\texttt{low}_{l,d}, \texttt{up}_{l,d}))$
13: $t \leftarrow h_d\,(e)$
14: **if** $t/2 < bdg - tot$ **then**
15: $r \leftarrow$ CREATERELATIONSHIP$(l, a, d, e, t)$
16: ADDRELATIONSHIP$(ego, r)$
17: $tot \leftarrow tot + t$
18: $i \leftarrow i + 1$
19: **else**
20: $done \leftarrow$ `True`
21: **end if**
22: **until** $done$
23: **return** $ego$ ▷ $s_{\text{net}}$ is the final value of $i$
24: **end procedure**

For each relationship, the algorithm has to set the type of the relationship $d$ and the gender of the alter $a$. The variable $d$ takes the values K and NK, in case of kin and non-kin relationship respectively. The variable $a$, such as $g$, takes values M (male) or F (female). The algorithm randomly extracts the values of $a$ and $d$ from the joint probability mass functions $f_{A,D}$. Since each layer has a different composition, which also depends on the gender of the ego, there is a specific function $f_{A,D|L=\bar{l},G=\bar{g}}$ for each layer $\bar{l}$ and for each gender $\bar{g}$. The functions refer only to the composition of the external part of the layers. Considering the current layer $l$ and the gender of the ego $g$, the algorithm extracts $a$ and $d$ from the function $f_{A,D|L=l,G=g}$ (line 11).

For each relationship, the algorithm has to assign a level of emotional closeness to the variable $e$ using the probability density functions $f_E$. There are two different $f_E$ functions, one to use in case of kin relationship $f_{E|D=\text{K}}$,

and the other for non-kin relationship $f_{E|D=\mathrm{NK}}$, therefore the algorithm selects the proper function according to $d$. The extraction from $f_{E|D=d}$ is limited in an interval of emotional closeness $\left(\mathtt{low}_{l,d}, \mathtt{up}_{l,d}\right)$ related to the current layer $l$ and the type of the current relationship $d$ (line 12). The method to infer these intervals is described in the Subsection IV.G.

To relate the emotional closeness $e$ to the time required to handle it, the algorithm is based on functions $h_d$ that return an amount of time given a level of emotional closeness. There are two different functions $h_{\mathrm{K}}$ and $h_{\mathrm{NK}}$, for kin and non-kin relationships respectively. Using the proper function $h_d$ the algorithm sets the amount of time $t$ given the level of emotional closeness $e$ (line 13). Functions $h_d$ must satisfy some properties listed in the Subsection IV.H.

The current relationship has to be added to the ego network only if there is enough time available. However if the algorithm discards a relationship when $t > bdg - tot$, the final value of $tot$ is always less than the time budget $bdg$. Since we want that $\mathbf{E}[tot] = \mathbf{E}[bdg]$ the condition to add a relationship to the network is $t/2 < bdg - tot$ (lines 14-18). When the previous condition gets false, the boolean control variable $done$ becomes equal to `True`, therefore the loop ends and the algorithm returns the object $ego$ with the related ego network (lines 19-23).

The final value of the counter $i$ represents the network size $s_{\mathrm{net}}$. If the functions and the parameters of the model are defined satisfying the properties given in the following subsections, the algorithm generates, on average, ego networks with the expected size $\mu_{\mathrm{net}}$.

## IV. Parameters and Functions

In this section we define all the parameters and functions the model uses exploiting results in the reference literature.

### A. Layer sizes

In the literature there are different values for the layer sizes, often with significant differences. In [7], the authors collected all the required data about layer sizes and extracted the mean value for each layer. Therefore, basing on this work, we set the mean support clique size $\mu_{\mathrm{sup}} = 4.6$, the mean sympathy group size $\mu_{\mathrm{sym}} = 14.3$ and the mean active network size $\mu_{\mathrm{net}} = 132.5$.

### B. Parameter $m$

Parameter $m$ is the probability to have a male ego, that is $gen = \mathrm{M}$. We can reasonably assume that $m = 0.50$.

### C. Function $f_S$

The sympathy group size distribution is presented in a histogram format ([8]) which can be fitted by a $Gamma$ distribution. As $f_S$ must be consistent with the mean size of the sympathy group $\mu_{\mathrm{sym}}$, we obtained $f_S = Gamma(4.1, 3.49)$ with mean 14.3.

TABLE I

COMPOSITION OF SYMPATHY GROUP

| $a$, $d$ | $g = \mathrm{M}$ | | $g = \mathrm{F}$ | |
|---|---|---|---|---|
| $a = \mathrm{M}$, $d = \mathrm{K}$ | 2.28 | 15.98% | 2.38 | 16.64% |
| $a = \mathrm{F}$, $d = \mathrm{K}$ | 2.47 | 17.26% | 3.53 | 24.72% |
| $a = \mathrm{M}$, $d = \mathrm{NK}$ | 7.38 | 51.61% | 2.02 | 14.14% |
| $a = \mathrm{F}$, $d = \mathrm{NK}$ | 2.17 | 15.15% | 6.36 | 44.51% |
| $sum$ | 14.3 | 100% | 14.3 | 100% |

### *D. Function $f_W$*

The ratio between the support clique and the sympathy group sizes is given by the function $f_W$. Since we have set the mean sizes $\mu_{\mathrm{sup}}$ and $\mu_{\mathrm{sym}}$, we define $f_W$ thought a normal distribution with mean equal to $\mu_{\mathrm{sup}}/\mu_{\mathrm{sym}} = 0.3217$. We have no explicit information about the standard variation of the distribution, however it can be experimentally approximated, using the scatter plot proposed in [8]. A good approximation is obtained by setting the standard variation to half of the mean, therefore the function is defined as $f_W = Normal(0.3217, 0.1608)$.

### *E. Function $f_B$*

We have no exact information about the distribution of time spent in socializing but we know that on average a person spends for it about the $20\%$ of the time [12]. Therefore we define $f_B$ with a mean value equal to $8760 \cdot 0.2 = 1752$ where $8760$ is the number of hours in a year. In this way expected value of time budget is $\mathbf{E}[bdg] = 1752$.

The probability function $f_B$ directly influences the distribution of the network sizes, therefore we chose its distribution and parameters experimentally, after we have done some tests, in order to obtain a network size distribution close to the one presented in [4]. The function we selected is $f_B = Gamma(205.48, 8.5264)$.

### *F. Functions $f_{A,D}$*

Dunbar & Spoors in [8] studied the composition of the sympathy group for male and female egos. Considering the given mean size $\mu_{\mathrm{sym}}$, that is independent of the gender of the ego, the resulting compositions are reported in Table I.

In the same work, the authors studied the support clique and they observed that there are not significant differences between the compositions of the two layers. For this reason we can set the function $f_{A,D|L=\mathrm{sym}}$, that refers to the external part of the layer, with the values in Table I, related to the whole sympathy group. Moreover we can set $f_{A,D|L=\mathrm{sup}} = f_{A,D|L=\mathrm{sym}}$.

Regarding the external part of the active network layer we can indirectly estimate its composition starting from results in [6]. Specifically, we set $f_{A,D|L=\mathrm{net}}$ with the results presented in Table II.

TABLE II
COMPOSITION OF ACTIVE NETWORK LAYER (EXTERNAL PART)

| $a$, $d$ | $g = \mathrm{M}$ | | $g = \mathrm{F}$ | |
|---|---|---|---|---|
| $a = \mathrm{M}$, $d = \mathrm{K}$ | 11.46 | 9.70% | 17.35 | 14.68% |
| $a = \mathrm{F}$, $d = \mathrm{K}$ | 18.00 | 15.23% | 17.18 | 14.53% |
| $a = \mathrm{M}$, $d = \mathrm{NK}$ | 52.50 | 44.41% | 38.90 | 32.91% |
| $a = \mathrm{F}$, $d = \mathrm{NK}$ | 36.24 | 30.66% | 44.78 | 37.88% |
| $sum$ | 118.2 | 100% | 118.2 | 100% |

### G. Emotional closeness intervals and functions $f_E$

As shown in Figure 2 the value of emotional closeness will be extracted by different range of the $f_E$ distribution depending on the layer. The intervals of emotional closeness can not be chosen arbitrarily but they must be consistent with the expected layer sizes ($\mu_{\mathrm{sup}}$, $\mu_{\mathrm{sym}}$ and $\mu_{\mathrm{net}}$) and with the probability density functions $f_E$. The probability to extract a value of emotional closeness in an interval must be equal to the proportion of the network the related layer represents.

Our model uses two different density functions for kin $f_{E|D=\mathrm{K}}$ and non-kin $f_{E|D=\mathrm{NK}}$, therefore, in order to define the intervals, we need to know the mean proportion of kin for each layer. Using the Equation (1) we obtain the probability $k'_l$ to have a kin in the external part of a layer $l$.

$$\begin{aligned} k'_l = \sum_{a \in \{\mathrm{M},\mathrm{F}\}} & \big(m \cdot f_{A,D|L=l,G=\mathrm{M}}\,(a,\mathrm{K}) \\ & + (1-m) \cdot f_{A,D|L=l,G=\mathrm{F}}\,(a,\mathrm{K})\big) \qquad ,\forall\, l \in \mathcal{L} \end{aligned} \tag{1}$$

Using the values $k'_l$ it is possible to obtain the probability to have a kin, $k_l$, in the whole layer $l$ by the Equation (2), where $c$ is a sublayer of $l$.

$$k_l = \sum_{c \subseteq l} \frac{\mu'_c}{\mu_l} \cdot k'_c \qquad ,\forall\, l \in \mathcal{L} \tag{2}$$

For example, the probability to have a kin in the whole network, $k_{\mathrm{net}}$, is:

$$k_{\mathrm{net}} = \frac{\mu'_{\mathrm{net}} \cdot k'_{\mathrm{net}} + \mu'_{\mathrm{sym}} \cdot k'_{\mathrm{sym}} + \mu'_{\mathrm{sup}} \cdot k'_{\mathrm{sup}}}{\mu_{\mathrm{net}}} \tag{3}$$

Considering a type of relationship $d$, the probability to extract a value from $f_{E|D=d}$ in the interval $\left(\mathtt{low}_{l,d}, \mathtt{up}_{l,d}\right)$ related to a layer $l$, must be equal to the expected proportion of the network the layer $l$ represents, considering only relationships with type $d$.

Knowing the cumulative distribution functions $F_E$ of the densities $f_E$, it is possible to calculate the limits of the

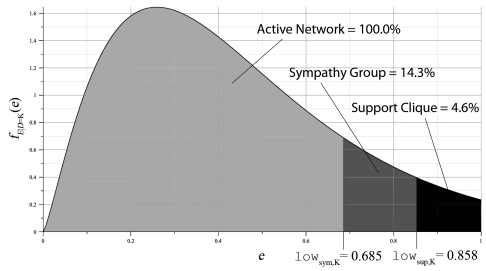


Fig. 2. Distribution of emotional closeness for kin with the proportion of the network for each layer and the related limits.

intervals of emotional closeness, considering them as quantiles that satisfy the following equations:

$$F_{E|D=\mathrm{K}}\left(\mathtt{low}_{\mathrm{sup,K}}\right) = 1 - \frac{\mu_{\mathrm{sup}} \cdot k_{\mathrm{sup}}}{\mu_{\mathrm{net}} \cdot k_{\mathrm{net}}} \tag{4}$$

$$F_{E|D=\mathrm{NK}}\left(\mathtt{low}_{\mathrm{sup,NK}}\right) = 1 - \frac{\mu_{\mathrm{sup}} \cdot (1 - k_{\mathrm{sup}})}{\mu_{\mathrm{net}} \cdot (1 - k_{\mathrm{net}})} \tag{5}$$

$$F_{E|D=\mathrm{K}}\left(\mathtt{low}_{\mathrm{sym,K}}\right) = 1 - \frac{\mu_{\mathrm{sym}} \cdot k_{\mathrm{sym}}}{\mu_{\mathrm{net}} \cdot k_{\mathrm{net}}} \tag{6}$$

$$F_{E|D=\mathrm{NK}}\left(\mathtt{low}_{\mathrm{sym,NK}}\right) = 1 - \frac{\mu_{\mathrm{sym}} \cdot (1 - k_{\mathrm{sym}})}{\mu_{\mathrm{net}} \cdot (1 - k_{\mathrm{net}})} \tag{7}$$

For example, considering kin relationships and the support clique layer, the limit $\mathtt{low}_{\mathrm{sup,K}}$ defines an area in $f_E$ whose size is equal to $\frac{\mu_{\mathrm{sup}} \cdot k_{\mathrm{sup}}}{\mu_{\mathrm{net}} \cdot k_{\mathrm{net}}}$ (the dark area in Figure 2) where $\mu_{\mathrm{sup}} \cdot k_{\mathrm{sup}}$ is the number of kin relationships in the support clique and $\mu_{\mathrm{net}} \cdot k_{\mathrm{net}}$ is the number of kin relationships in the whole network.

The lower limits for the active network layer are $\mathtt{low}_{\mathrm{net},d} = 0$ while the upper limits are $\mathtt{up}_{\mathrm{sup},d} = e_{max}$, where $e_{max}$ is the max value of emotional closeness, $\mathtt{up}_{\mathrm{sym},d} = \mathtt{low}_{\mathrm{sup},d}$ and $\mathtt{up}_{\mathrm{net},d} = \mathtt{low}_{\mathrm{sym},d}$, for each type of relationship $d$.

Distributions of emotional closeness for kin and non-kin are presented in [6]. As we do not have the exact distributions' values, we can only approximate them. Setting the maximum level of emotional closeness $e_{max} = 1$, obtained distributions are $f_{E|D=\mathrm{K}} = Gamma(0.2, 2.296)$ and $f_{E|D=\mathrm{NK}} = Normal(0.5, 0.172)$, both defined only in the interval $(0, e_{max})$. Considering the cumulative distributions $F_E$ it is possible to solve the Equation (4), (5), (6) and (7), obtaining the limits of the intervals of emotional closeness: $\mathtt{low}_{\mathrm{sup,K}} = 0.8582$, $\mathtt{low}_{\mathrm{sup,NK}} = 0.8185$, $\mathtt{low}_{\mathrm{sym,K}} = 0.6852$ and $\mathtt{low}_{\mathrm{sym,NK}} = 0.7247$.

### *H. Functions $h_d$*

$h_d$ functions correlates the level of emotional closeness to the time spent in a relationship. Considering the studies [4] and [5] we know that $h_d$ functions are increasing with the level of emotional closeness and that $h_K$ returns lower or equal values than $h_{NK}$. The latter observation is due to the fact that kin relationships demand less time invested on them than non-kin relationships. However, for high level of emotional closeness, the invested time in social relationships is equal for both kin and non-kin, therefore we set the following constraint:

$$h_{\mathrm{K}}\left(e_{max}\right) = h_{\mathrm{NK}}\left(e_{max}\right) \tag{8}$$

TABLE III
RESULTS: LAYER SIZES AND TIME BUDGET

| | min | max | avg | st. dev. |
|---|---|---|---|---|
| $s_{\text{net}}$ | 3 | 510 | 132.84 | 65.80 |
| $s_{\text{sym}}$ | 0 | 74 | 14.06 | 7.25 |
| $s_{\text{sup}}$ | 0 | 43 | 4.62 | 3.55 |
| *bdg* | 195.62 | 5197.87 | 1748.40 | 598.42 |

where $e_{max}$ is the maximum level of emotional closeness.

Since the network size $s_{\text{net}}$ is limited by time constraints, it is fundamental to properly define the functions $h_d$ in order that $\mathbf{E}[s_{\text{net}}] = \mu_{\text{net}}$. In order to do this we impose that, in an average network with size $\mu_{\text{net}}$, the total amount of time spent in relationships is equal to the main value of the time budget $\mathbf{E}[s_{bdg}]$, obtained from the density function $f_B$. Considering the given density functions of emotional closeness $f_E$ and the proportion of kin in the network $k_{\text{net}}$, the constraint can be expressed by the Equation (9). In this equation, the value of the integral is the weighted sum of the expected values of the functions $h_{\text{K}}$ and $h_{\text{NK}}$, multiplied for the probability to have a kin or a non-kin respectively.

$$\mu_{\text{net}} \cdot \int \big[h_{\text{K}}(e) \cdot f_{E|D=\text{K}}(e) \cdot k_{\text{net}} + h_{\text{NK}}(e) \cdot f_{E|D=\text{NK}}(e) \cdot (1 - k_{\text{net}})\big] \, \mathrm{d}e = \mathbf{E}[bdg] \tag{9}$$

Through the graphics in [5] and in [4], we presume that $h_d$ functions have an exponential trend therefore we define a generic $h$ function: $h(e) = c^e + t_0 - 1$. The parameter $t_0$ is the value returned by $h(0)$. It can be considered as the minimum amount of time spent in a relationship in order to keep it active.

$h_{\text{K}}$ and $h_{\text{NK}}$ have the same form as $h$ but have different values for the parameters $c$ and $t_0$: respectively $c_{\text{K}}$ and $t_{0\text{K}}$ in $h_{\text{K}}$, and $c_{\text{NK}}$ and $t_{0\text{NK}}$ in $h_{\text{NK}}$.

As previously said $h_K$ has to return lower or equal values than $h_{NK}$ therefore $t_{0\text{K}}$ must be less or equal than $t_{0\text{NK}}$. We have no any indication on how estimate $t_0$ parameters, therefore we assume to be reasonable to set $t_{0\text{K}} = 0.5$ and $t_{0\text{K}} = 2$. In order to extract parameters $c$ we can put in a system the Equation (8) and (9) where $\mu_{\text{net}} = 132.5$, $k_{\text{net}} = 0.2817$ and $\mathbf{E}[bdg] = 1752$.

With numeric methods we can solve the system of equations with a very good approximation obtaining $c_{\text{K}} = 95.3275$ and $c_{\text{NK}} = 93.8275$. Finally we can define the functions:

$$h_{\text{K}}(e) = 95.3275^e - 0.5 \tag{10}$$

$$h_{\text{NK}}(e) = 93.8275^e + 1 \tag{11}$$

## V. RESULTS

We have implemented the algorithm presented in Section III.A in Java programming language and we performed 100.000 run tests creating as many ego network graphs. Results are presented in the Tables III and IV.

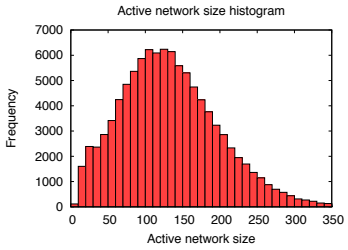


Fig. 3. Network Sizes Distribution in Simulations

TABLE IV

RESULTS: COMPOSITION OF THE NETWORK

| $a_i$, $d_i$ | $g = \mathrm{M}$ (49.85%) | | $g = \mathrm{F}$ (50.15%) | |
|---|---|---|---|---|
| $a_i = \mathrm{M}$, $d_i = \mathrm{K}$ | 13.63 | 10.35% | 19.97 | 14.89% |
| $a_i = \mathrm{F}$, $d_i = \mathrm{K}$ | 20.37 | 15.48% | 20.93 | 15.61% |
| $a_i = \mathrm{M}$, $d_i = \mathrm{NK}$ | 59.44 | 45.17% | 41.50 | 30.96% |
| $a_i = \mathrm{F}$, $d_i = \mathrm{NK}$ | 38.16 | 29.00% | 51.68 | 38.54% |
| $sum$ | 131.59 | 100% | 134.08 | 100% |

As we can see, the average network size converges to a value close to the expected value $132.5$. The little gap is due to approximation errors.

Also the mean average of the sympathy group is very close to the reference value $14.3$. In this case the gap is due to the correlation between the time budget and the size of the layer. The algorithm extracts $s_{\mathrm{sym}}$ values from the distribution $f_S$ but in a few cases the algorithm exhausts the time budget before completing to populate the sympathy group layer, making lower its mean size. This happens especially when the algorithm extracts a low value for $bdg$. In our tests, the sympathy group size is constrained by time budget in the $3.17\%$ of the runs.

The average size of the support clique meets perfectively its expected value. Such as in case of the sympathy group, the time budget extracted can constrain the size of the layer however, in case of the support clique, this happened only in the $0.38\%$ of the runs.

The shapes of the layer size distributions are similar to the distributions in the reference literature. See for example the shape of the network size distribution presented in Figure 3.

In Table IV we can see that the composition of the network is coherent with the $f_{A,D}$ functions we set. Male egos have smaller network than females. This is due to female egos have a little more kin relationships which request less time that non-kin relationships.

We have validated the model, demonstrating that it allows generating ego network graphs that are coherent with the results in the reference literature.

## VI. Extended Social Networks

The ego network model, defined in section III, allows us to generate typical social structures from the ego's point of view, however it does not provide any information about how different ego networks are connected to each other. In order to develop and validate protocols for data dissemination in the cyber-physical world, researchers need to have at their disposal the whole social network graph. The next step of our work is, hence, to extend our model in order to allow the generation of wide social network graphs, which have to be coherent with the defined ego network model and with available real data.

In the next subsection we presents a method for the generation of extended social network graphs. This is a preliminary work that currently lacks in a deep analysis and in a comparison with available real data for different configurations, however obtained results satisfy some common hypothesis about social networks, like the small-world property.

### *A. Assumptions*

In social networks, links are typically reciprocal: if a person has an active social relationship with another person, there is a high probability to have an active relationship also in the opposite direction. This is especially true for strong relationships. In any case reciprocal social ties can have different strength. In this preliminary work we consider that every link is bidirectional and it has the same strength in both direction. In future work, the model will be extended in order to consider also unidirectional ties.

Defining the procedure for the generation of social network graphs we introduced some simplifications in respect to the ego network model. The gender of the nodes is not taken into account and the social relationships are not divided into kin and non-kin ones. Also in this case, the model may be extended in future in order to consider different kinds of egos and relationships.

### *B. Overview*

An extended social network can be defined as the social graph resulting from the interconnection of the nodes' ego networks. For this reason the idea is to initialize the network with a certain amount of nodes and then execute for each node the generative procedure defined for the ego network model.

Nodes are initialized assigning the attributes related to their ego networks: the time budget, the support clique and the sympathy group sizes. Ego networks are generated at the same time, starting from the inner toward the outer layers, like in the ego network model. At each step, the procedure selects a pair of nodes between which create a new relationship, until all the ego networks are completed.

Once the algorithm selected the first of two nodes involved in the new relationship, the choice of the other end-point is fundamental in order to determine the social structure of the network. We know that social networks exhibit the *small-world property* therefore nodes are organized in communities where interconnections are more common than in random graphs. A simple and effective way to obtain this property is to induce the *triadic closure*.

The method consists of two steps: first, a node *k* is chosen from the neighbors of the selected node *i*, then a link is created between *i* and a node *j*, chosen from the neighbors of *k* apart from *i*.

As Granovetter's studies demonstrate [13], the triadic closure is more probable if the strength of the existing links between the nodes *i* and *k* and between *k* and *j* is high. This is intuitively true for real human social networks. For this reason our procedure selects with a higher probability the nodes which have stronger ties with the starting node.

However in real human network, not all the relationships born because of a common friends. Sometime a new relationship is created with a node of the network regardless the current ego network structure, choosing the node apparently in a random fashion.

Our procedure for the generation of social networks takes into account both methods to form a new link: *triadic closure method* (TC) and *random selection method* (RS).

We can reasonably suppose that the inner layers of different ego networks are highly clusterized among each other therefore, in our procedure, TC method has to be selected with higher probability for inner layers than for the outer ones. For this reason our procedure relies on parameters $p_{\mathrm{sup}}$, $p_{\mathrm{sym}}$ and $p_{\mathrm{net}}$ which define the probability to select TC method against RS method for each different layer.

### C. The Algorithm

The algorithm we are going to present, generates a social network following the procedure introduced in the Section VI-B.

In the first part of the algorithm (lines 2-6), data structures and objects are initialized, taking the parameter $N$ as the number of nodes to be created.

Each node is created by the procedure CREATEEGO (line 4) which also initializes its ego network parameters: the size of the sympathy group $s_{\mathrm{sym}}$, the size of the support clique $s_{\mathrm{sup}}$ and the time budget $bdg$. Each parameter is extracted from the proper distribution function as in the ego network model.

The sets $A_l$ contain the nodes those have available resources to form new relationships for the layer $l$, that is the nodes having free slots, in case $l \in \{\mathrm{sup}, \mathrm{sym}\}$, or those having enough residual time budget, in case $l = \mathrm{net}$. The algorithm initializes these sets with the whole set of nodes $V$ (line 6).

The main loop (lines 7-29) manages the construction of the network from the inner layer ($\mathrm{sup}$) to the outer layer ($\mathrm{net}$). For each layer $l$, the algorithm generates relationships until the set of nodes with available resources $A_l$ is empty (lines 8-28).

At each step of the inner loop, the algorithm tries to create a new relationship between a node $i$ and a node $j$. The node $i$ is chosen from the set $A_l$ proportionally to its available resources for the current layer $l$ (line 9). Therefore node $i$ is selected with probability $slot_l[i] / \sum_{i^\star \in A_l} slot_l[i^\star]$, where $slot_l$ is the number of slots available for the current layer $l$, in case of $l \in \{\mathrm{sup}, \mathrm{sym}\}$, and with probability $rds[i] / \sum_{i^\star \in A_l} rsd[i^\star]$, where $rsd$ is the residual time budget ($bdg - tot$), in case of $l = \mathrm{net}$.

**Algorithm 2** Create Social Network

**procedure** CREATENETWORK($N$)
  initialize sets of nodes $V$ and edges $E$
  **for** $i \leftarrow 1, N$ **do**
    $V[i] \leftarrow$ CREATEEGO
  **end for**
  $A_{\text{sup}}, A_{\text{sym}}, A_{\text{net}} \leftarrow V$
  **for all** layer $l$ in $\{\text{sup}, \text{sym}, \text{net}\}$ **do**
    **while** $A_l$ is not $empty$ **do**
      $i \leftarrow$ WEIGHTEDSELECTION($A_l$)
      $e \leftarrow$ EXTRACTFROM($f_E \; in \; (\texttt{low}_l, \texttt{up}_l)$)
      $t \leftarrow h(e)$
      **if** $t/2 < bdg[i] - tot[i]$ **then**
        $j \leftarrow null$
        **if** RAND() $< p_l$ **then**
          $j \leftarrow \text{TC}(V, A_l, i, t)$
        **end if**
        **if** $j$ is $null$ **then**
          $j \leftarrow \text{RS}(V, A_l, i, t)$
        **end if**
        **if** $j$ is not $null$ **then**
          $E \leftarrow E +$ CREATELINK($i, j, e$)
          UPDATE($i$, $j$, $A_l$)
        **else**
          RECOVERY
        **end if**
      **else**
        $A_l \leftarrow A_l - i$
      **end if**
    **end while**
  **end for**
  **return** $V$, $E$
**end procedure**

TABLE V
MEASURES OF GENERATED SOCIAL NETWORK

| | |
|---|---|
| Average degree | 132.55 |
| Clustering coefficient | 0.0695 |
| Assortativity | 0.0002 |
| Average path length | 2.2488 |
| Support clique average size | 4.59 |
| Sympathy group average size | 14.07 |

Emotional closeness of the new relationship $e$ is extracted from $f_E$ distribution function, within the bounds $\mathrm{low}_l$ and $\mathrm{up}_l$ (line 10), while the time required $t$ is calculated with the function $h$ (line 11). Like in the ego network model, the algorithm checks if the selected node $i$ has enough residual time budget for create the new relationship which requires time $t$ (line 12). If it is not, node $i$ is removed from the set $A_l$ (line 26).

As introduced in the Section VI-B, node $j$ can be selected through TC or RS method. TC method is chosen with probability $p_l$, while RS method is chosen with probability $p_l - 1$ or if TC method fails (lines 16-18).

Let $K$ the set of the neighbors of the selected node $i$, TC method selects a node $k$ in $K$ with probability $e_{ik} / \sum_{k^\star \in K} e_{ik^\star}$, where $e_{ik}$ is the emotional closeness level of the existing relationship between $i$ and $k$. Once selected the node $k$ and let $J$ the set of its neighbors, the procedure try to extract a node $j$ from the set $J^\star = J \cap A_l - K - \{i\}$ with probability $e_{kj} / \sum_{j^\star \in J^\star} e_{kj^\star}$. After the selection of the node $j$, the procedure checks if selected node has enough residual time budget for the new relationship requiring time $t$. If it is not, node $j$ is removed from the set $A_l$ and TC method tries to select another node $j$ given the node $k$. If it is not possible because of $J^\star$ is empty, the procedure restarts selecting another node $k$. If for each node $k$ selected, it is not possible to select a node $j$, the TC procedure fails.

In RS method, the selection of the node $j$ is performed in a pure random fashion from the set $A_l - \{i\}$. Also in this case the algorithm checks if the selected node has enough residual time budget for the new relationship. If it is not, it is removed from the set $A_l$. If $A_l = \{i\}$, the RS procedure fails.

If the algorithm fails selecting a node $j$, a procedure recoveries the deadlock adding resources for a random node $j$ or reducing them for the node $i$ (line 24).

On the contrary, if nodes $i$ and $j$ are selected, the algorithm creates a new relationship, with emotional closeness level $e$, between them (line 21). In the next step, UPDATE procedure calculates the new values for $rsd$ and $slot_l$ for the nodes $i$ and $j$. Then, according to these values, the set $A_l$ is updated, eventually removing the nodes, if they exhausted the available resources for the current layer $l$ (lines 22).

### *D. Simulations*

The model of social network presented in this section shares many parameters with the ego network model introduced in Section III. The functions $f_S$, $f_W$ and $f_B$, used to initialize the nodes in the network, are defined in

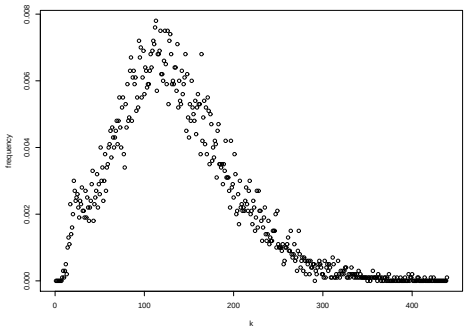


Fig. 4. Degree Distribution in Simulated Social Network

the same manner as in the Subsections III-C, III-D and III-E. Removing the categories kin and non-kin we need to define a new function $f_E$ and, consequently, new layer limits and a new $h$ function.

In these preliminary simulations we set the function $f_E$ as a Normal distribution with mean 0.54 and variance 0.25. The limits of the emotional closeness intervals, calculated with the same procedure as in the Subsection III-G are $\mathrm{low}_{\mathrm{sup}} = 0.9174$ and $\mathrm{low}_{\mathrm{sym}} = 0.8161$ while $h$ function is defined as $h(e) = 62.27^e + 1$.

We performed a preliminary simulation with $N = 10000$, $p_{\mathrm{sup}} = 1.0$, $p_{\mathrm{sym}} = 0.9$ and $p_{\mathrm{sup}} = 0.6$. Relevant measures of the generated social network graph are reported in Table V.

As we can expect, the ego network properties, like the average degree and the average layer sizes, are coherent with the results presented in Section IV. Also the obtained distribution of the degree in Fig. 4 is almost identical to the distribution reported in Fig. 3.

Clustering coefficient is calculated as the average of the local clustering coefficients defined in [14]. In respect to the clustering coefficients reported by Newman for some real social networks in [15], generated network exhibits a lower value. Further analysis will demonstrate if this gap is acceptable. An analysis of the relation between the clustering coefficient of a node and its degree is reported in Fig. 5. It has a very similar trend compared with the same analysis performed on a virtual social network in [16].

The average path length index is very low, demonstrating that generated network has the small-world property.

Assortativity index indicates that there is no correlation between the degrees of connected nodes. However social networks reported by Newman in [15] always exhibits a certain degree of positive assortativity. Also in this case further studies are necessary in order to know if our values is acceptable or not.

We also analyzed the similarity between connected nodes, defined as the number of neighbors they share. In Fig. 6 we analyze the relation between the number of common neighbors and the emotional closeness level of the relationship. As we can expect, nodes with strong relationships share a higher number of friends, on average about 17. Moreover we can observe remarkable discontinuities between different layers.

## ACKNOWLEDGMENT

This work was funded by the European Commission under the FIRE SCAMPI (258414) project. Authors wish to acknowledge the very fruitful discussions with Prof. R. I. M. Dunbar of the University of Oxford, which have

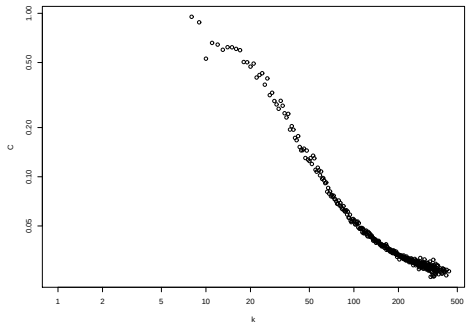


Fig. 5. The relation between the local clustering coefficient $C$ of the degree $k$.

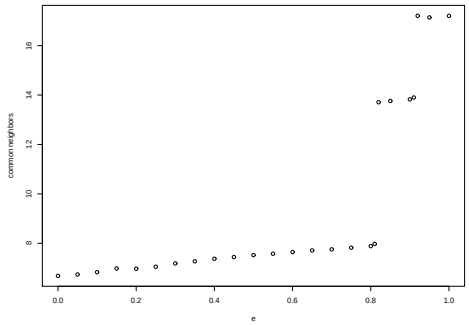


Fig. 6. The relation between the number of common neighbors of each pair of connected nodes and the level of emotional closeness $e$ of the relationship.

been fundamental to deeply understand the structures of human social networks, and how to correctly model them.

## REFERENCES

[1] M. Conti *et al.*, "Looking ahead in pervasive computing: challenges and opportunities in the era of cyber-physical convergence," *Pervasive and Mobile Computing Journal*, 2011, in press.

[2] T. V. Pollet, S. G. B. Roberts, and R. I. M. Dunbar, "Use of social network sites and instant messaging does not lead to increased social network size, or to emotionally closer relationships with offline network members," *Cyberpsychology, Behavior, and Social Networking*, 2010, in press.

[3] R. I. M. Dunbar, "The social brain hypothesis," *Evolutionary Anthropology: Issues, News, and Reviews*, vol. 6, no. 5, pp. 178–190, 1998.

[4] R. A. Hill and R. I. M. Dunbar, "Social network size in humans," *Human Nature*, vol. 14, no. 1, pp. 53–72, 2003.

[5] S. G. B. Roberts and R. I. M. Dunbar, "Communication in social networks: Effects of kinship, network size, and emotional closeness," *Personal Relationships*, 2010, in press.

[6] S. G. B. Roberts *et al.*, "Exploring variation in active network size: Constraints and ego characteristics," *Social Networks*, vol. 31, no. 2, pp. 138–146, May 2009.

[7] W. X. Zhou *et al.*, "Discrete hierarchical organization of social group sizes," *Proceedings of the Royal Society B: Biological Sciences*, vol. 272, no. 1561, pp. 439–444, 2005.

[8] R. I. M. Dunbar and M. Spoors, "Social networks, support cliques, and kinship," *Human Nature*, vol. 6, no. 3, pp. 273–290, Sep. 1995.

[9] J. Stiller and R. I. M. Dunbar, "Perspective-taking and memory capacity predict social network size," *Social Networks*, vol. 29, no. 1, pp. 93–104, Jan. 2007.

[10] R. S. Burt, "Decay functions," *Social Networks*, vol. 22, no. 1, pp. 1–28, May 2000.

[11] L. Schuchman, "Dither signals and their effect on quantization noise," *Communication Technology, IEEE Transactions on*, vol. 12, no. 4, pp. 162–165, Dec. 1964.

[12] R. I. M. Dunbar, "Theory of mind and the evolution of language," in *Approaches to the Evolution of Language: Social and Cognitive Bases*, J. R. Hurford, M. Studdert-Kennedy, and C. Knight, Eds. Cambridge: Cambridge University Press, 1998.

[13] M. Granovetter, "The Strength of Weak Ties," *The American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[14] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998. [Online]. Available: http://dx.doi.org/10.1038/30918

[15] M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[16] A. Grabowski, "Interpersonal interactions and human dynamics in a large social network," *Physica A-Statistical Mechanics and its Applications*, vol. 385, no. 1, pp. 363–369, NOV 1 2007, pT: J; PG: 7.