Measuring the Collective Potential of Populations from Dynamic Social Interaction Data

Manuel Cebrian*, Mayank Lahiri, Nuria Oliver, and Alex (Sandy) Pentland

***Abstract*—In any society, is the way in which individuals interact, intentionally or unintentionally, designed to maximize global benefit, or does it result in a fundamentally non-egalitarian stratification of society, where a small number of individuals inevitably dominate? Our ability to observe and record interactions between individuals in real populations has improved dramatically with modern technological improvements, but it is still a difficult task to use this data to model cooperation and collaboration between individuals, and its global effect on the entire population. To shed light on these questions, we model an individual's value in society as an epistatic mathematical function of a set of binary choices, and the *collective potential* of a population as the expected value of an individual over time. Individuals try to selfishly improve their societal value by adopting the choices of their neighbors, constrained by the actual observed interaction topology and order. As a result, we are also able to investigate how far natural populations are from an optimal regime of functioning. We show that interaction topology has a large impact on collective potential, but the relative order of specific interactions seems to have a negligible effect.**

***Index Terms*—Social networks, social factors, genetic algorithms**

I. Introduction

THE study of how people interact in successful corporations is providing managers with better tools to allocate human resources, organize work, and be more efficient in general [1]–[3]. Similarly, natural science research into social insect behavior is helping computer scientists design robust distributed control and optimization algorithms [4]–[6]. Cooperation and competition within natural populations are the bedrock of complex social structures, but although our technological ability to observe the dynamics of interactions within these populations has improved, modeling and quantifying the global sociological effects of these dynamics remains a difficult problem. There has also been little work in quantifying how far a population of entities is from its *optimal* regime of functioning in terms of global wellness of the entire population, and what little literature does exist is not intended for social interpretations [7].

This problem is fundamentally connected to an important question in social science that concerns the interplay between *individual* and *collective* success in social networks: how does a person's interactions with other people affect their social position? Furthermore, how is society at large globally influenced by the collective effects of these local interactions [8]–[10]? In the social sciences, this has long been the focus of a 'positivist' line of thinking, which defines social progress as the changing of society towards an ideal state, generated by individual contributions and aggregated by collective interactions [11], [12]. It is also the question we investigate in this paper, albeit in a strictly quantitative sense.

We leverage the unprecedented opportunities offered by the recent availability of large amounts of social interaction data, such as email and phone call records, and ideas from optimization theory and social network analysis, to analyze real populations in terms of the questions just posed. The result is a novel framework that allows us to characterize populations from social interaction data in a mathematically robust way, based on the population's intrinsic ability for local interactions to produce a positive global outcome over time. We describe this as the *collective potential* of a population, and analyze several real networks in terms of the impact of interaction topology and order on its collective potential.

Our framework for computing the collective potential of a population is based on the notion of a hypothetical *collective potential curve*. Each individual in a population has a dynamic state at any given time, which we model as a real-valued function of multiple, interacting binary choices that the individual has made. The interactions between these binary choices, and thus the overall state function, may be made *arbitrarily complex*, with the end result being that the 'value' of each individual within the population is expressed as a single, continuous value.

M.C. (cebrian@media.mit.edu) and A.P. (sandy@media.mit.edu) are with The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139.

M.L. (mlahir2@uic.edu) is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607. Parts of the current research were performed while M.L. and M.C. were at Telefonica Research in Madrid, Spain.

N.O. (nuriao@tid.es) is with Telefonica Research, Barcelona, Spain.

We also assume that each individual seeks to increase their state value by interacting with their neighbors over time, and adopting some of their neighbors' more beneficial choices, *i.e.,* when imitating the neighbor's choices would result in the selfish positive outcome of increased state value.

The collective potential curve of a population is then defined as the trajectory of the expected state value in the population over time[1], with a key contribution of our method being that the expectation is computed over all possible state functions in a computationally tractable way. The collective potential curve therefore represents the efficiency of constructive, collaborative processes in a population over time, or how efficiently the structure and dynamics of social interactions can foster positive global change in the population through selfish local interactions. Although an expectation over all possible state functions cannot be computed analytically, we show that in practice, computational simulation estimates of the collective potential curves of large, real populations converge very quickly, usually in a few dozen iterations, even with populations of millions of individuals.

The collective potential curve presents some interesting avenues for the analysis of populations from dynamic social interaction data. Diffusion processes that take place in social networks have been studied in the context of epidemiological modeling [13], [14], and more recently in 'viral marketing' scenarios, where word-of-mouth recommendations drive the adoption of a product [15], [16]. Since the collective potential curve models a form of diffusion through a dynamic network over time, it allows us to explicitly compare the effect of interaction order and topology on the *efficiency* and *speed* of diffusion, and also to compare the dynamics of different populations which have been controlled for size and other external factors.

Our definition of the collective potential curve of a population has its roots in a number of research areas. The global dynamics of individuals following binary choice models has been studied extensively in mathematical sociology [17]–[19]. There has also been interest in how the structure of the network impacts the rates of diffusion of information [20], [21]. Our contribution here is to model a population by means of an arbitrarily complex state function that operates on a multitude of choices made by each individual. We use a simple model of interactions between individuals – determined by real interaction data[2] – and vary the complexity of the state function, computing an expectation over all possible functions in a countably infinite class of state functions [22]. The sociological idea of a society progressing towards an ideal state through interactions between individuals is modeled by a form of *collective optimization* [23], [24]. The specific type of collective optimization we use is a modified form of a simple genetic algorithm [25], which bears similarities to parallel genetic algorithms with spatially distributed populations and mating topologies [26]–[29]. The generation of arbitrarily complex state functions is based on a class of synthetic functions called *Hyperplane Defined Functions* (HDFs), initially devised as difficult test cases for genetic optimization methods [22].

This paper is organized as follows. In the next section, we describe our framework in detail. In particular, we describe the Simple Genetic Algorithm (SGA) as an optimization technique, which is central to our framework, as well as the nature of Hyperplane-Defined Functions (HDFs) as objective functions of arbitrary complexity.

In Section III we present a detailed experimental study to evaluate our ideas on a number of real dynamic networks and summarize our findings. Finally, our conclusions and lines of future work are presented in Section IV.

II. Collective Potential in Dynamic Networks

Our framework quantifies a number of sociological principles in as simple a way as possible. Each individual is represented by a binary state vector that encodes a set of choices it has currently made, without specifying the form or function of each choice. By allowing the state vectors to grow arbitrarily long, we can encode any number of choices. A global objective function operates on the state vectors and assigns each one an objective score, as a measure of value for its choices. Although the presumption of a global measure of worth for all individuals might violate some sociological principles, we compensate by allowing the objective function to be arbitrarily complex.

Individuals in the population seek to increase their own worth, which they achieve by interacting with other individuals. We assume a simple model of interactions between individuals, where the topology and order of interactions between individuals are governed by recorded data. For example, in a dataset of phone call records, an

[1] Note that other statistics of the distribution of state values are equally applicable.

[2] It should be noted that social interaction data is often easier to collect than actual diffusion data, which is why stochastic diffusion models are used to estimate the efficiency and extent of diffusion.

undirected or mutual interaction between two individuals takes place when they call each other. Similarly, if the dataset consists of email records, an email sent between two addresses qualifies as a *directed* interaction from the sender to the recipient, *i.e.*, the sender gains no advantage in sending the email, but the recipient might. Furthermore, communications networks are by no means the only type of data that can be used. Interaction networks in the recent past have been derived from physical proximity determined by BlueTooth sensing devices [30], [31], wearable badges [2], radio tracking collars on wild animals [32], and bibliographic databases of co-publication patterns [33], among others.

During each interaction, a random subset of choices (state) is temporarily exchanged between the pair of interacting individuals. This exchange becomes permanent if the value of the state of either individual increases (unless the interactions are directed, in which case only the recipient's state can change). Although this is a very simple model of interactions between individuals, it is surprisingly flexible when paired with an appropriate objective function[3]. Furthermore, by holding the interaction model constant, the only variable in our model is the objective function.

The collective potential curve of the population is defined as the rate of increase in the expectation of the objective value of individuals in a population over time. The expectation is computed over all possible objective functions, which encompass a large class of collective behavior models. Thus, our method is essentially parameter free. This allows us to measure how effective a population is at spreading positive processes, which may be choices, ideas, information, or any of a number of other diffusion processes.

In the next subsection, we describe the interaction algorithm in detail, and in the following subsection, the choice of objective function.

A. Genetic Optimization in Dynamic Networks

In the most general form, dynamic networks consist of a set of individuals $V = \{v_1, ..., v_n\}$ interacting with each other over a period of T discrete timesteps. The exact definition of what constitutes an individual or interaction depends on the domain being analyzed, and is not central to our problem. Let $\mathbf{x}_i^{(t)}$ be a binary string of length L representing the state of individual v_i at time t, randomly initialized according to some distribution. Let $f(\mathbf{x})$ be an objective function (described in detail in Section II-B) that assigns an objective score to any state string:

$$f(\mathbf{x}) : \{0,1\}^L \mapsto \mathbb{R}$$

We assume that exactly one interaction $e_t = (v_i, v_j)$ occurs at each time step t, although this assumption can be easily relaxed. The interaction model is similar to the formulation of a simple genetic algorithm [25], [34], [35], and is formally defined as follows.

1) Let v_i and v_j be the two individuals interacting at time t, with corresponding state strings $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$. A *crossover point* c is selected uniformly at random from the integer range $[1, L]$.
2) Two new state strings are created by swapping the tails of $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$, where the tail is defined as all positions including and after index c. Let these two new state strings be $\mathbf{y}_1$ and $\mathbf{y}_2$. When $c = 1$ or $c = L$, crossover has no effect, since $\mathbf{y}_1$ and $\mathbf{y}_2$ are necessarily identical to $\mathbf{x}_j^{(t)}$ and $\mathbf{x}_i^{(t)}$ respectively. Consider the following example:

$$\mathbf{x}_i^{(t)} = \langle 1,1,1,1,1,1 \rangle$$
$$\mathbf{x}_j^{(t)} = \langle 0,0,0,0,0,0 \rangle$$

If $c = 4$, we would get the following two new strings after crossover:

$$\mathbf{y}_1 = \langle 1,1,1,0,0,0 \rangle$$
$$\mathbf{y}_2 = \langle 0,0,0,1,1,1 \rangle$$

3) The objective score of each new state string is then evaluated. If any of them have a greater objective score than either of their parents $\mathbf{x}_i^{(t)}$ or $\mathbf{x}_j^{(t)}$, the corresponding parent's state string is replaced for the next iteration.

$$\mathbf{x}_i^{(t+1)} = \underset{x \in \{\mathbf{x}_i^{(t)}, \mathbf{y}_1, \mathbf{y}_2\}}{\arg\max} \; f(x)$$
$$\mathbf{x}_j^{(t+1)} = \underset{x \in \{\mathbf{x}_j^{(t)}, \mathbf{y}_1, \mathbf{y}_2\}}{\arg\max} \; f(x)$$

In the case of ties in the objective scores of the original and a new string, the original state string is retained. This is an important step if the resultant model is to serve as a general case of certain network diffusion models in the literature, but is not required in general.
4) The steps above are repeated for all recorded interactions in increasing time order. After each interaction, the average objective score in the population is measured.

The process described in steps 1–4 is repeated for multiple random trials, with different random initial states, until the average objective score curve computed

[3] It can be shown that many social network diffusion models in the literature can be reduced to collective potential processes. This is subject of ongoing research and will be published elsewhere.

in step 4 stabilizes. We show in Section III that these estimates of the expectation converge very quickly, usually in less than a few dozen random trials. Note that the procedure outlined above is similar to collective optimization [23], parallel or spatial genetic algorithm with distributed populations [24], [26]–[29], and the economic analysis of interacting agents distributed in a networks [24], [36]. To the best of our knowledge, our research is the first to work towards a convergence of these topics, bridging the gap between microdynamics of individual level success and macrodynamics of global success.

B. Hyperplane-Defined Functions

A standard tool for the analysis of canonical genetic algorithms is Holland's Schema Theorem [35]. Along with the building block hypothesis proposed as a refinement of it in [25], it explains the ability of a GA to solve optimization problems by manipulating short binary substrings occurring at specific positions, called *schemata*[4], that contribute to an increase in the objective value of a longer string that contains them. The crossover operator can then be seen as a mechanism for probabilistically exchanging schemata between a set of random strings. As strings of lower fitness are replaced with strings of higher fitness, the average and maximum objective values of the population increase over time, eventually leading to the discovery of a local or global optimum. Although other mechanisms have been proposed to analyze canonical genetic algorithms[5], as well as a proliferation of algorithmic variants (see, for example, [39] and [34]), the use of schemata is appropriate in our context.

In particular, we are concerned with a class of schemata-based, synthetic objective functions called *hyperplane-defined functions* (HDFs), which were originally designed to serve as difficult benchmark functions for assessing the performance of different genetic algorithms [22]. An HDF is constructed by selecting a set of schemata that contribute a certain value to the overall objective score of the string. The schemata are chosen randomly and hierarchically, starting with relatively short schemata of order 1 occurring at random starting positions within the string. Pairs of such schemata are concatenated to generate schemata of order 2, and so on, with each schema receiving an individual positive or negative score. The end result is an objective function that takes a binary string as input and returns an objective score that is the sum of the scores of all the individual schemata contained in it. Further details are described in Holland [22], which shows how HDFs can be used to generate objective functions of arbitrary optimization difficulty (*e.g.*, nonlinear, discontinuous, nonseparable, nonsymmetric functions).

The following example illustrates the generation of a simple HDF $f(\mathbf{x}) \mapsto \mathbb{R}$ that takes a binary string of length $n = 10$ as input. The asterisk character ('*') denotes a 'don't care' character that matches any binary value. The following are the set of randomly chosen schemata that define the HDF.

`*01*******`	*score 2, order 1 schema*
`*****11***`	*score 2, order 1 schema*
`********10`	*score 3, order 1 schema*
`*01**11***`	*score -4, order 2 schema*
`*01*****10`	*score 4, order 2 schema*

The following binary strings are now evaluated using the HDF above:

`0010000000`	*score:* 2
`0010011000`	*score:* $2+2+(-4)=0$
`1010011010`	*score:* $2+2+3+(-4)+4=7$

As noted earlier, HDFs are an infinite class of functions of arbitrary complexity, and are thus appropriate for modeling a number of complex phenomena. They are a natural representation for information exchange, where each schema represents a unit of information, and different units can interact in complex ways. A large number of local minima, plateaus, discontinuities, and other difficult optimization landscape features ensure that it is not trivial to optimize an HDF. In our experiments, we use the method described in [22] to construct HDFs – including the artificial truncation of negative objective values to zero that yields a non-negative objective function – and generate two higher orders of schemata.

III. Experimental Results

We now present an experimental study of our technique applied to four real dynamic network datasets. Our primary objective is to analyze the shape of the collective potential curves for each dataset, and how these shapes change under different kinds of random perturbations of the original interactions. We are also concerned with how quickly collective potential curves stabilize to their final shape, and an analysis of the role

[4] *sing.* schema

[5] There is some debate about whether the Schema Theorem explains the genetic algorithm's ability for efficient optimization [22], [34], [37], [38], but the specific aspects of the Schema Theorem being debated do not affect our use of schemata here.

Dataset	Individuals	Interactions	Type	Real time frame
CDR-J	$\sim 10^6$	$\sim 10^7$	Undirected	$\sim$ 4 months
CDR-C	$\sim 10^5$	$\sim 10^7$	Undirected	$\sim$ 6 months
Enron Emails	84,716	1,343,655	Directed	4 years, 3 months (September 1998 - December 2002)
Peer-to-peer	5,955	81,297	Directed	45 days (April 2003 - May 2003)

TABLE I
DATASET CHARACTERISTICS

that specific individuals play in the process. Each facet of the analysis will be dealt with in a separate section below, following a description of the datasets we used.

A. Datasets

We used four real dynamic network datasets for our experimental evaluation, two of which are publicly available. Table I lists the characteristics of each dataset. Due to privacy and confidentiality considerations, we only state orders of magnitude for the *CDR-J* and *CDR-C* datasets.

1) **(CDR-J)** Telecommunications companies collect Call Detail Records (CDRs) from telephone subscribers for a variety of reasons, including billing and network performance analysis. These records generally contain details of every call attempted, made, received, and dropped, as well as other information like talk time. We obtained fully anonymized, one-way encrypted CDR data from a large telecommunications provider for a random sample of customers from a fixed geographical area. Each individual is a telephone subscriber, and an interaction occurs between two individuals when one makes a phone call to the other. We ignore the direction of the call because phone conversations (as opposed to phone *calls*) are inherently bi-directional.
2) **(CDR-C)** This dataset is similar to the *CDR-J* dataset, except that the random sample of customers is chosen from a completely different geographical area. Like the other CDR dataset, *CDR-C* is completely anonymized and one-way encrypted.
3) **(Enron Emails)** As part of its investigation into corporate fraud at the now defunct Enron Corporation, the United States Federal Energy Regulatory Commission obtained and publicly released the complete email records of 150 (former) Enron executives[6]. The dataset contains records of all messages sent, received, and deleted by each of the executives, resulting in a very rich email dataset. We used a version of the dataset with 84,716 email addresses and 1,343,655 unique, non-duplicate emails. Each individual in the network is an email address, and a directed interaction occurs when an email is sent from one address to another.
4) **(Peer-to-peer file sharing)** We use publicly released, anonymized traces of peer-to-peer file sharing in an internal university network [40]. Each individual is a unique computer, and a directed interaction occurs between two computers when one downloads a file from the other. Although the original study states that 6,528 users shared 291,925 files over 81 days, we were able to parse far fewer file transfers from the publicly available data, as shown in Table I. Note that although the files transferred were usually audio files, there was no such restriction in place, and the transfer of arbitrary files through peer-to-peer networks is a common vector for computer virus propagation.

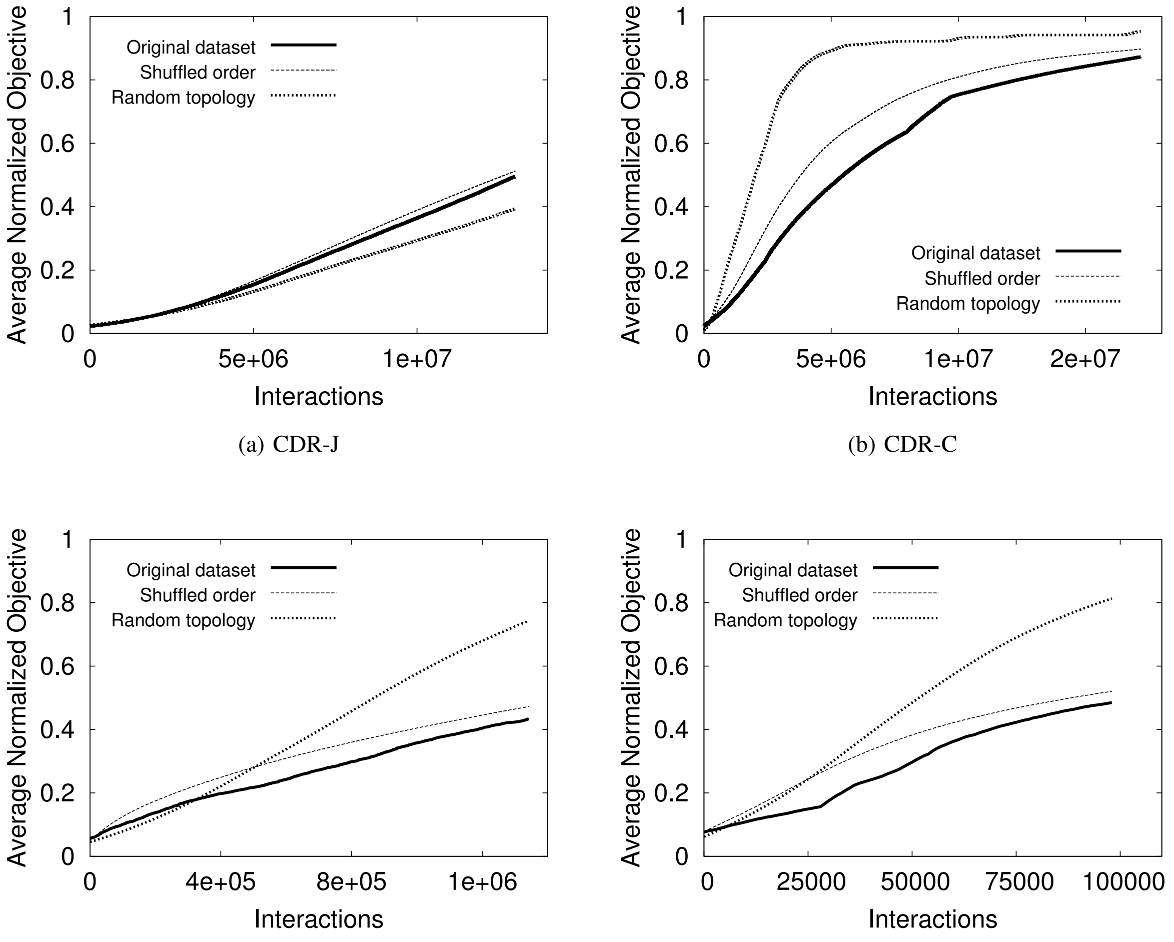
B. Methodology

We ran 200 random trials for each dataset, each with a different randomly generated HDF objective function. The HDFs were chosen to operate on state vectors of a random dimension $n \in [128, 2048]$ bits. State vectors were randomly initialized. Note that in some situations, it might be useful to guarantee that initial state vectors are somehow uniformly of 'low' fitness. However, since the express purpose behind the use of an HDF is to generate a difficult optimization function, it is hard to determine the global optimum (or optima), and therefore to determine what constitutes a state vector of low fitness.

We ran three different sets of experiments with the same number of individuals and interactions: one with the original interaction data as given by the experimental datasets, another where the interacting individuals were chosen randomly (*random topology*)[7], and a third with

[6] Publicly available at http://www.cs.cmu.edu/~enron/

[7] This represents a random dynamic interaction topology, and is equivalent to sequentially sampling edges uniformly at random from a complete graph.



(c) Enron Emails

(d) Peer-to-peer

Fig. 1. Collective potential curves for the original datasets as well as their perturbed versions.

the original interactions taking place in a shuffled order (*shuffled interaction order*). These two additional experiments allow us to analyze the impact of topology and interaction order on the collective potential curves. We also ran a similar set of experiments using randomly generated HDFs operating on state vectors of the same dimension, with the difference in results being negligible.

C. Collective Potential Curves

The first aspect of our analysis is to investigate the shape of collective potential curves for each of the datasets described in Section III-A. Figure 1 shows these curves for all four datasets. What can we expect these curves to look like? Recall that the collective potential curve is the expectation of the average objective value in the population over time, computed over all possible objective functions. Since HDFs are artificially truncated to non-negative values, the curve will be positive for all points. Since individuals only change their choices (state) if it is beneficial to them, the curve will also be non-decreasing. Furthermore, we are not interested in the actual objective values of the function, but rather the average value relative to the maximum value discovered at the end of each run. Thus, the curves are normalized by the maximum discovered objective value (hence, their values are in the [0, 1] interval). The number steps of the simulation equals the number of interactions recorder for each dataset.

Impact of Topology: Our first observation is that the shape of the collective potential curve for random topologies matches what is theoretically expected from random mixing. Giacobini *et al.* [29] analyzed the average fitness of individuals in GAs with spatially structured populations, including randomly structured populations, and our empirical results are in agreement with theirs. Note that although epidemics (and thus general diffusion processes) are expected to spread quickly in scale-free networks [41], [42], a random topology has the advantage (or disadvantage, depending on the scenario) of bridging the gap between isolated communities or graph components in the long run.

Although the shape of the random topology curve in the *CDR-J* dataset seems to be qualitatively different from the other datasets, this may be explained by the large size of the population relative to the number of recorded interactions – *CDR-C*, in comparison, has an equivalent number of interactions, but its population is an order of magnitude smaller, being much denser. Thus, the curve presented in Figure 1(a) is comparable to the early stages of the curves in other datasets. As the number of interactions grows, a random interaction topology represents the best possible scenario for collective potential in the long run, since it is equivalent to randomly sampling edges sequentially and uniformly at random, with replacement, from a complete (clique-like) interaction topology. Thus, isolated communities are broken down, similar to the spread of ‘good’ schemata in parallel genetic algorithms [27]. Note that this is the asymptotic behavior of the random topology as the number of interactions grows, not the short-term behavior over a possibly small, fixed number of interactions.

Impact of Interaction Order: We also note that shuffling the order of interactions seems to always have a positive impact on the collective potential curve, *i.e.*, the average objective value of the population increases faster than with the original order of interactions. There is a very simple, plausible explanation for this phenomena. It is well understood that many types of real networks exhibit a high degree of clustering [43], [44], and that the composition of clusters in networks change over time [45], [46]. A good set of schemata might therefore get ‘stuck’ by circulating within a cluster, unable to spread to other nodes in the population outside the cluster, until nodes that carry the good schemata change affiliations to a different cluster. This is a well known notion in evolutionary biology and ecology [47] and, closer to our context, in the study of parallel genetic algorithms with distributed populations [27]. By shuffling the order of interactions in the original dataset, we help to break up some of these clusters in the time dimension, thereby allowing better mixing of good schemata throughout the entire population.

It is important no note that shuffling has a larger positive impact on CDR-J than on CDR-C (subfigures 1(b) and 1(b)). Although there are many plausible reasons for this phenomenon, one possibility is that the geographical region covered by the CDR-J is known to be much larger and also have a higher level of regional deprivation than the region covered by CDR-C, a facet which has been previously connected to a lower communication diversity using phone datasets [30]. Shuffling the order of the interactions breaks down temporal clusters and increases the level of communication diversity on average over time, which would then manifest in a more pronounced lift in the collective potential curve than in CDR-C.

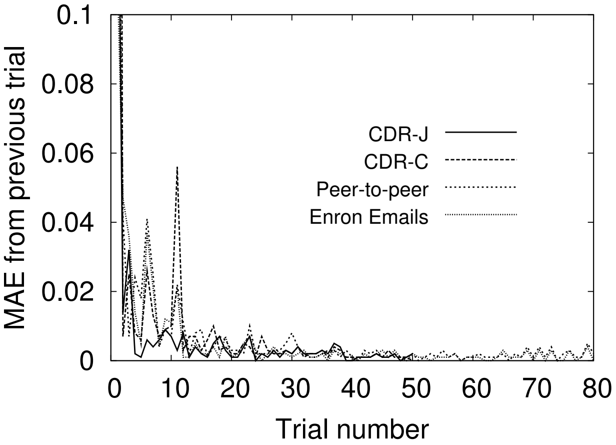


Fig. 2. Convergence of collective potential curves for the original datasets.

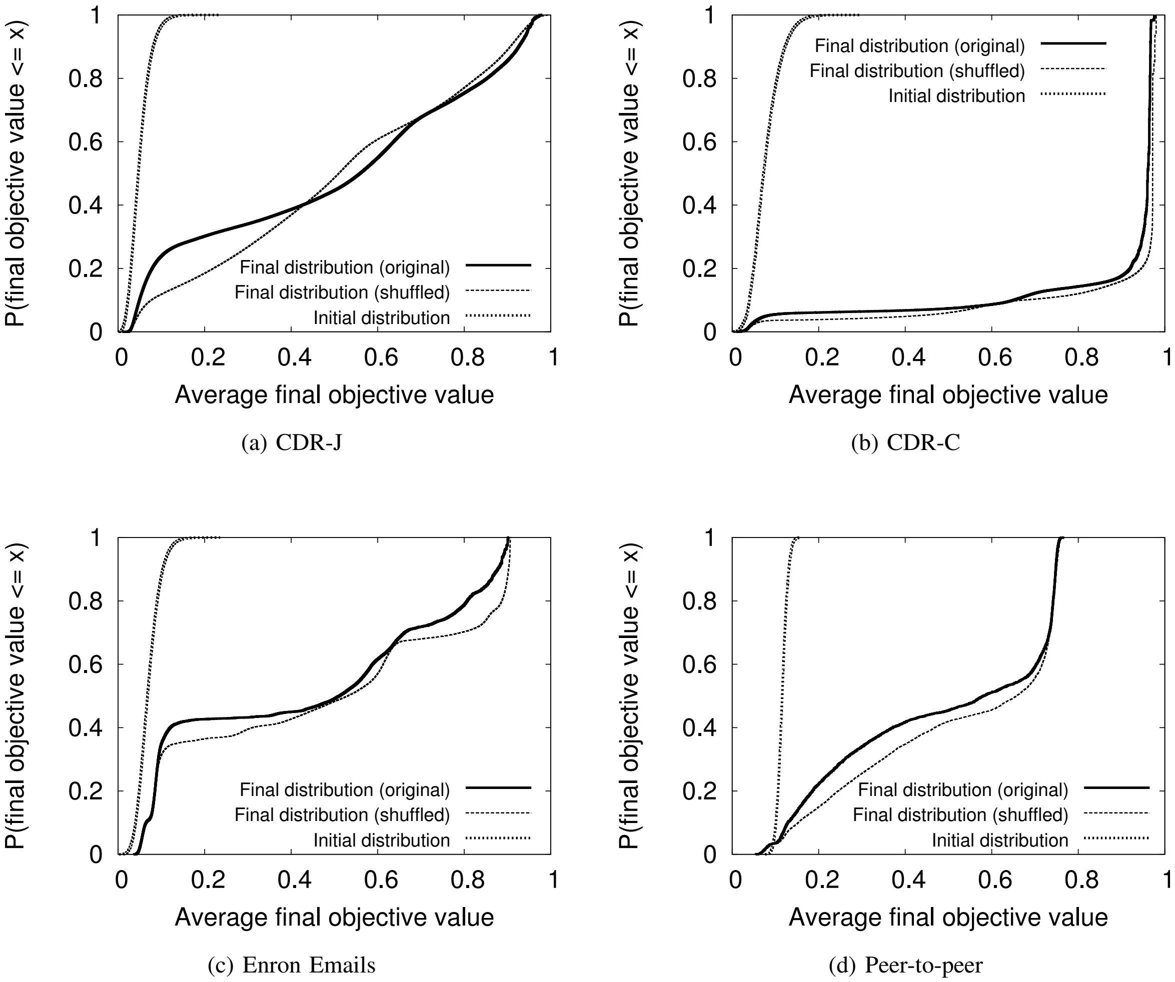
D. Convergence Analysis

The collective potential curves in Figure 1 are produced by stochastic processes, so a natural question is to ask how stable they are, or how many random trials are required till they converge to a robust estimate. For each trial, we calculate the average deviation of the curve from the average curve computed up to the previous iteration. Figure 2 shows this value at each trial for the original dataset curve in Figure 1, where the average absolute deviation is used as the distance measure between successive curves and the average up to the previous trial.

The results indicate that the collective potential curves settle to their expectation very quickly, usually requiring less than 15 trials for the mean absolute deviation to drop to less than 1%. We obtain similar results using different distance measures, such as the root-mean-squared deviation and the maximum deviation, which indicate that the collective potential curves are quite stable and tractable to compute, even for large datasets.

E. Node Effects

In the previous two sections, we analyzed the global collective potential of our experimental populations. We now turn our attention to individuals within the



(a) CDR-J

(b) CDR-C

(c) Enron Emails

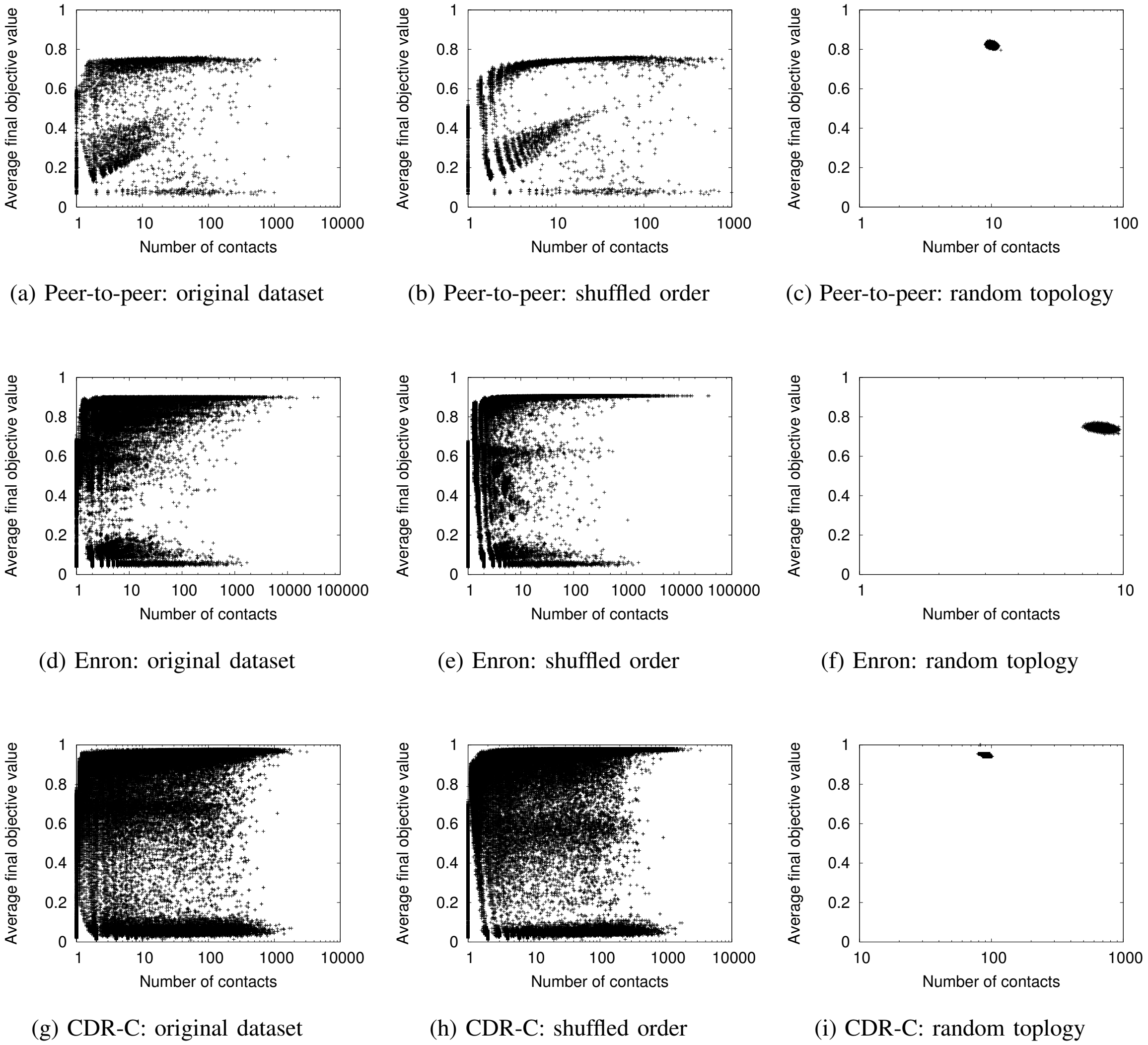
(d) Peer-to-peer

Fig. 3. Empirical cumulative distributions for the average final objective values of nodes in each population.

populations, particularly to the distribution of objective values in the population at the end of processing all interactions. Recall that the initial state vectors were randomly chosen binary strings. Figure 3 shows the cumulative distribution of objective values for all four datasets at the end of processing all interactions, averaged over 200 random trials.

Final objective distributions: The empirical distributions of final objective values show some common features. Perhaps the most prominent feature is a knee in the lower x-axis of the distribution. This is particularly pronounced in the *CDR-J* (Figure 3(a)) and Enron (Figure 3(c)) datasets, at final objective values of $x \sim 0.05$ and $x \sim 0.1$ respectively. These knees represent a significant section of the population that does not advance in fitness values relative to the top performer; approximately 20% of nodes for *CDR-J* and 40% for the Enron dataset are in this region. Similar knees are evident in the other two datasets, representing isolated islands of low objective values.

Although the reason for this disparity could be intrinsic structural differences in the populations, another possible reason is that it is an artifact of the data collection process. For the CDR datasets as well as the Enron dataset, a number of phone numbers and email addresses are included in the dataset simply as a result of communicating with a set of 'core' nodes. In the CDR examples, these core nodes are customers of the telecommunications company who are included in the random sample, and in the Enron dataset, the core consists of executives whose mailboxes were subpoenaed, as well as their close associates. The core nodes are unique because all or most of their interactions are actually observed. In these datasets, there is by extension a significant 'periphery' of nodes, whose activity is not completely observed. As a result, isolated from the active core, they do not get the same opportunity to interact



(a) Peer-to-peer: original dataset

(b) Peer-to-peer: shuffled order

(c) Peer-to-peer: random topology

(d) Enron: original dataset

(e) Enron: shuffled order

(f) Enron: random toplogy

(g) CDR-C: original dataset

(h) CDR-C: shuffled order

(i) CDR-C: random toplogy

Fig. 4. Average final objective value of each individual compared to the number of contacts (interactions) it is involved in.

and increase in their objective values as the core nodes. This is fundamentally an issue of missing data, and is an important and largely unsolved consideration in all of networks research.

However, it is interesting to note there is no such issue with the peer-to-peer dataset. All computers involved in the file sharing study were directly observed as part of the experiment, so there is no core or periphery. Yet, we see similar features in the final fitness distributions, which could indicate that at least some part of the features in the other distributions are intrinsic to the structure and dynamics of the populations.

Stratification of Societies: The final objective value distributions shown in Figure 3 raise interesting questions about which characteristics of the datasets cause stratified distributions in final objective values. As mentioned earlier, there is generally a core of nodes for whom we have complete information, with the remaining nodes being discovered incidentally in the course of interactions with core nodes. We can therefore expect core nodes to have many more observed interactions than periphery nodes, which could result in their having more chances to increase their state, and thus a higher final objective value on average than periphery nodes.

Figure 4 shows scatter plots of the average final objective value of each individual plotted against the number of interactions it is involved in, and Figure 5 plots the same data against the number of neighbors for each node (unique interactions). Despite the fact that we are not able to find a strong trend, there is certainly some structure in the result. On the one hand, having a *large number of interactions* and even more importantly, having a *large neighborhood* leads to high final fitness, as seen in both figures. On the other hand, having a high final fitness does not necessarily imply a large number of interactions or a large neighborhood. This is true for those individuals who are possibly locked inside an isolated island of low fitness, with no interactions that can reach the population at large. The frequency of their interactions (Figure 4) is irrelevant, given that they always talk to a set of isolated individuals of low average fitness. Barring other contributing factors, a possible sociological explanation of this is the following: if you live in a very isolated town, it does not matter if you talk to a few or many people, as you have few connections to the outside world.

If the stratification observed in the final objective distributions were an artifact of missing data, then we would expect a trend in the average final fitness as a function of the number of contacts an individual is involved in. However, no such trend is visible in either the original datasets (Figures 4(a), 4(d), and 4(g)) or their shuffled versions (Figures 4(b), 4(e), and 4(h)). We do not plot the *CDR-J* dataset for practical purposes, as it would contain more than 10^6 data points.

It should be noted that the random topology serves as a model case of an egalitarian society. Individuals are assigned initial state vectors from the same (uniform) random distribution, and chosen randomly for interactions with other individuals. The topology and order of interactions is also random, which results in the contact frequencies of all individuals being drawn from the same multinomial distribution. As a result, in the random topologies, individual objective values and contact frequencies are distributed in a relatively small range, as is theoretically expected. There is relatively little differentiation between the individuals with the highest and lowest final objective values in a population where everyone gets the same opportunities for interaction.

IV. Conclusions and Future Work

In this paper, we have defined and computed collective potential curves based on real social interaction data by means of a robust, parameter-free, estimate of the capacity of a population to increase their collective wellness in a given time period. Next, we have empirically investigated the impact that the network topology and the order of interactions have on the collective potential of a population. Our results are compatible with known results from population genetics and evolutionary computation, namely that networks with random topology asymptotically yield the highest collective potential, and small levels of perturbation in the timing of the interactions help to prevent inbreeding of good solutions, and tend to speed-up the collective potential growth rate. Finally, it is interesting to note that under our model, having a large number of contacts with other individuals and a large neighborhood of contacts leads to a high final fitness, but the converse is not necessarily true, *i.e.,* certain individuals consistently achieve a high final fitness value in spite of having smaller and more infrequent circles of contacts.

Our experimental results also show a number of interesting features that warrant further sociological analysis and experiments. For example, shuffling the order of interactions in the CDR-J dataset has a more pronounced effect on the lift in collective potential that shuffling the CDR-C dataset. It is know that the region the CDR-J dataset was drawn from has a higher level of regional deprivation than CDR-C. Since shuffling the interaction order breaks down temporal clusters, it is interesting to ask whether sociological factors such as an inherently low level of communication diversity in the CDR-J region is responsible for this phenomenon.

In future work, we envision two directions for our research. The first deals with validating and explaining the sociological implications of our findings here. Some possible questions are to ask if there are commonalities in the local network properties of individuals at each strata of society, or if the collective potential curves are correlated with global network properties such as hierarchy or community structure. A second line of research would be to investigate the nature of the genetic stochastic process we have described in this paper, in terms of its capabilities as a collective behavior model, its convergence, and other possible uses.

References

[1] A. Pentland, *Honest signals: how they shape our world.* The MIT Press, 2008.

[2] L. Wu, B. Waber, S. Aral, E. Brynjolfsson, and A. Pentland, “Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task,” in *International Conference on Information Systems*, 2008.

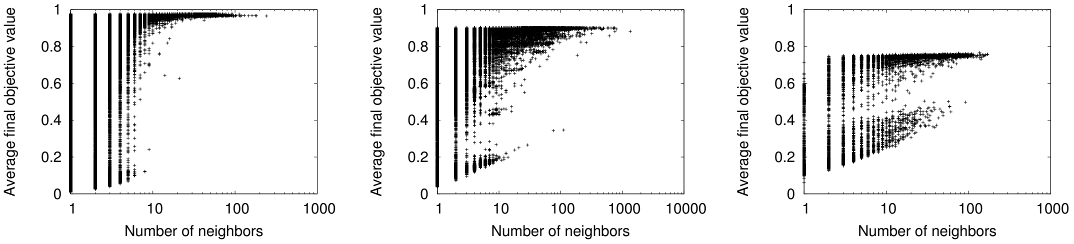
[3] D. Olguin Olguin, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, “Sensible organizations: Technology and methodology for automatically measuring organizational behavior,” *IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics*, 2009.

[4] E. Bonabeau, M. Dorigo, and G. Theraulaz, “Inspiration for optimization from social insect behaviour,” *Nature*, vol. 406, no. 6791, pp. 39–42, 2000.

[5] H. Maier, A. Simpson, A. Zecchin, W. Foong, K. Phang, H. Seah, and C. Tan, “Ant colony optimization for design of water distribution systems,” *Journal of Water Resources Planning and Management*, vol. 129, p. 200, 2003.

[6] V. Cicirello and S. Smith, “Wasp-like agents for distributed factory coordination,” *Autonomous Agents and Multi-agent systems*, vol. 8, no. 3, pp. 237–266, 2004.

[7] O. Kinouchi and M. Copelli, “Optimal dynamical range of excitable networks at criticality,” *Nature Physics*, vol. 2, no. 5, pp. 348–351, 2006.



(a) CDR-C (b) Enron (c) Peer-to-peer

Fig. 5. Neighbors

[8] G. Mugny and W. Doise, “Socio-cognitive conflict and structure of individual and collective performances,” *European Journal of Social Psychology*, vol. 8, no. 2, 1978.

[9] R. Slavin, “Research on cooperative learning and achievement: What we know, what we need to know,” *Contemporary educational psychology*, vol. 21, no. 1, pp. 43–69, 1996.

[10] J. Tudge, “Processes and consequences of peer collaboration: A Vygotskian analysis,” *Child Development*, pp. 1364–1379, 1992.

[11] A. Comte and J. Bridges, *A general view of positivism*. Kessinger Publishing, 2006.

[12] H. Spencer, *The study of sociology*. Adamant Media Corporation, 2000.

[13] R. M. May and A. L. Lloyd, “Infection dynamics on scale-free networks,” *Phys. Rev. E*, vol. 64, no. 6, p. 066112, Nov 2001.

[14] M. Newman, “Spread of epidemic disease on networks,” *Physical Review E*, vol. 66, no. 1, p. 16128, 2002.

[15] P. Domingos and M. Richardson, “Mining the network value of customers,” in *KDD ’01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2001, pp. 57–66.

[16] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in *KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 137–146.

[17] W. Brock and S. Durlauf, “Identification of binary choice models with social interactions,” *Journal of Econometrics*, vol. 140, no. 1, pp. 52–75, 2007.

[18] D. Lopez-Pintado and D. Watts, “Social influence, binary decisions and collective dynamics,” *Rationality and Society*, vol. 20, no. 4, p. 399, 2008.

[19] G. Fischer, E. Giaccardi, H. Eden, M. Sugimoto, and Y. Ye, “Beyond binary choices: Integrating individual and social creativity,” *International Journal of Human-Computer Studies*, vol. 63, no. 4-5, pp. 482–512, 2005.

[20] D. Centola and M. Macy, “Complex contagions and the weakness of long ties 1,” *American Journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007.

[21] Y.-E. Lu, S. Roberts, P. Lio, R. Dunbar, and J. Crowcroft, “Size matters: Variation in personal network size, personality and effect on information transmission,” *IEEE International Conference on Computational Science and Engineering*, vol. 4, pp. 188–193, 2009.

[22] J. H. Holland, “Building blocks, cohort genetic algorithms, and hyperplane-defined functions,” *Evolutionary computation*, vol. 8, no. 4, pp. 373–391, 2000.

[23] L. Scardovi and R. Sepulchre, “Collective optimization over average quantities,” in *Proceedings of the 45th IEEE Conference on Decision and Control*, 2006, pp. 3369–3374.

[24] F. Stonedahl, W. Rand, and U. Wilensky, “Multi-agent learning with a distributed genetic algorithm,” in *AAMAS08: ALAMAS+ALAg Workshop*, 2008.

[25] D. E. Goldberg, *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[26] Y. Min, X. Jin, X. Su, and B. Peng, “Empirical analysis of the spatial genetic algorithm on small-world networks,” *Lecture Notes in Computer Science*, vol. 3993, p. 1032, 2006.

[27] E. Cantu-Paz, “A survey of parallel genetic algorithms,” *Calculateurs Paralleles, Reseaux et Systems Repartis*, vol. 10, no. 2, pp. 141–171, 1998.

[28] J. Payne and M. Eppstein, “Emergent mating topologies in spatially structured genetic algorithms,” in *Proceedings of the 8th annual conference on genetic and evolutionary computation*. ACM New York, NY, USA, 2006, pp. 207–214.

[29] M. Giacobini, M. Tomassini, and A. Tettamanzi, “Takeover time curves in random and small-world structured populations,” in *GECCO ’05: Proceedings of the 2005 conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2005, pp. 1333–1340.

[30] A. Pentland, “Reality mining of mobile communications: Toward a new deal on data,” *The Global Information Technology Report – World Economic Forum*, vol. 1, no. 6, pp. 79–80, 2009.

[31] N. Eagle and A. Pentland, “Reality mining: sensing complex social systems,” *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[32] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. I. Rubenstein, “Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet,” *ACM SIGPLAN Notices*, vol. 37, no. 10, pp. 96–107, 2002.

[33] A. L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, “Evolution of the social network of scientific collaborations,” *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3–4, pp. 590–614, 2002.

[34] M. D. Vose, *The simple genetic algorithm: foundations and theory*. MIT press, 1999.

[35] J. H. Holland, *Adaptation in natural and artificial systems*. Cambridge, MA, USA: MIT Press, 1992.

[36] R. Cowan and N. Jonard, “Network structure and the diffusion of knowledge,” *Journal of Economic Dynamics and Control*, vol. 28, no. 8, pp. 1557–1575, 2004.

[37] L. Altenberg, “The schema theorem and Prices theorem,” *Foundations of genetic algorithms*, vol. 3, pp. 23–49, 1995.

[38] G. Rudolph, “Convergence analysis of canonical genetic algorithms,” *IEEE transactions on neural networks*, vol. 5, no. 1, pp. 96–101, 1994.

[39] S. Baluja and R. Caruana, “Removing the genetics from the standard genetic algorithm,” in *ICML*, 1995, pp. 38–46.

[40] A. Fast, D. Jensen, and B. N. Levine, “Creating social networks to improve peer-to-peer networking,” in *KDD ’05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. New York, NY, USA: ACM, 2005, pp. 568–573.

[41] M. Shirley and S. Rushton, “The impacts of network topology on disease spread,” *Ecological Complexity*, vol. 2, no. 3, pp. 287–299, 2005.

[42] R. Pastor-Satorras and A. Vespignani, “Epidemic dynamics in finite size scale-free networks,” *Physical Review E*, vol. 65, no. 3, p. 35108, 2002.

[43] M. McPherson, L. Smith-Lovin, and J. Cook, “Birds of a feather: Homophily in social networks,” *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[44] M. Newman, “Clustering and preferential attachment in growing networks,” *Physical Review E*, vol. 64, no. 2, p. 25102, 2001.

[45] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, “A framework for community identification in dynamic social networks,” in *KDD ’07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2007, pp. 717–726.

[46] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, “Community evolution in dynamic multi-mode networks,” in *KDD ’08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2008, pp. 677–685.

[47] D. Charlesworth and B. Charlesworth, “Inbreeding depression and its evolutionary consequences,” *Annual Review of Ecology and Systematics*, vol. 18, no. 1, pp. 237–268, 1987.

Manuel Cebrian is a Postdoctoral Fellow with the Media Laboratory at MIT (Cambridge, MA, USA). Before joining MIT he was a Research Scientist at Telefonica Research (Madrid, Spain). Prior to Telefonica Dr. Cebrian was a Postdoctoral Fellow with the Department of Computer Science at Brown University (Providence, RI, USA). Dr. Cebrian holds a Ph.D. in Electrical Engineering and Computer Science from Autonomous University of Madrid (Madrid, Spain).

Mayank Lahiri is currently a Ph.D. candidate in the Department of Computer Science at the University of Illinois at Chicago. His interests are in dynamic social network analysis, machine learning, data mining, and computational population biology when it involves field work with large and dangerous animals.

Nuria Oliver is currently the Scientific Director for the Multimedia and Data Mining & User Modeling Research Teams in Telefonica Research (Barcelona, Spain). She received the BSc (honors) and MSc degrees in Electrical Engineering and Computer Science from the ETSIT at the Universidad Politecnica of Madrid (UPM), Spain, in 1992 and 1994 respectively. She received her PhD degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in June 2000. From July 2000 until November 2007, she was a researcher at Microsoft Research in Redmond, WA. At the end of 2007, she returned to Spain to create and lead the Multimedia Scientific Team at Telefonica Research in Barcelona. Since March 2009, she is also the acting Scientific Director for the Data Mining & User Modeling Team in Telefonica Research. Her research interests include mobile computing, multimedia data analysis, search and retrieval, smart environments, context awareness, statistical machine learning and data mining, artificial intelligence, health monitoring, social network analysis, computational social sciences, and human computer interaction. She is currently working on the previous disciplines to build human-centric intelligent systems.

Alex Pentland MIT Professor Alex (Sandy) Pentland is a pioneer in organizational engineering, mobile information systems, and computational social science. Sandy’s focus is the development of human-centered technology, and the creation of ventures that take this technology into the real world.

He directs the Human Dynamics Lab, helping companies to become more productive and creative through organizational engineering, and the Media Lab Entrepreneurship Program, which helps translate cutting-edge technology into real-world impact around the world. He is among the most-cited computer scientists in the world.