

PhD THESIS

**Time Evolution and Predictability
of Social Behavior in Techno-Social
Networks**

Antonia Godoy Lorite

Copyright © 2015 All Rights Reserved

UNIVERSIDAD ROVIRA I VIRGILI

PhD THESIS

**Time Evolution and Predictability
of Social Behavior in Techno-Social
Networks**

Author

Antonia Godoy Lorite

Supervisors

Marta Sales-Pardo

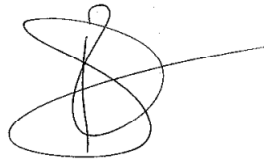
Roger Guimerà

DEPARTMENT OF CHEMICAL ENGINEERING

Tarragona, 2015

WE STATE that the present study, entitled “Time Evolution and Predictability of Social Behavior in Techno-Social Networks”, presented by Antonia Godoy Lorite for the award of the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of this university, and that it fulfils all the requirements to be eligible for the International Doctorate Award.

Doctoral Thesis Supervisor/s

A handwritten signature in black ink, consisting of a large, stylized 'R' followed by a horizontal line extending to the right.

Dr. Roger Guimerà Manrique

A handwritten signature in black ink, consisting of a cursive 'M' followed by a stylized 'S' and a vertical line.

Dra. Marta Sales Pardo

Tarragona, 30th November 2015

Agradecimientos

Llegando el fin de estos cuatro años que han sido mi doctorado, me gustaría dar las gracias a todas las personas que lo han hecho posible.

Primeramente, quisiera dar las gracias a mis supervisores Roger y Marta por muchas cosas, por haberme enseñado tanto, por su esfuerzo y dedicación, por haberse preocupado de formar investigadores en todo lo que conlleva, y en definitiva por ser un ejemplo de bien hacer a nivel profesional y humano.

Muchísimas gracias a todos los seeslabers, tantos los que quedan como los que estuvieron. A Núria, Arnau, Francesco, Manu, Toni Aguilar, Toni Valles, Oriol, Marc y Pedro. Con vosotros todo ha sido mucho más sencillo y agradable. Es un privilegio levantarte por la mañana con ganas de ir al lab (entendedme, ya una vez despierta, un poco despejada y desayunada, sólo entonces) y eso es gracias a vosotros. Habéis sido el apoyo práctico y moral ante las dificultades, y hemos compartido muchos muy buenos momentos, que espero seguir compartiendo siempre. También gracias al lab hermano (que no enemigo) del Suscape, en especial a Janire, por tantísimas risas que hemos echado todos juntos cada día dentro y fuera de la universidad. Sois todos fantásticos.

Muchas gracias a Dr. Christopher Moore, por acogerme en el Santa Fe institute, enseñarme tantas cosas y cuidar de que sacara el máximo de la estancia. Y a Laura Ware, porque ella fue mi gran compañera en ese tiempo, enseñándome lo que es la vida americana, incluyéndome como uno más de la familia. No hubiera sido lo mismo sin ella.

Gracias también a mis inseparables María, Teresa y Uri, por haber estado siempre, aun cuando es físicamente imposible. Ellos saben lo que ha costado esta tesis, y sé que también se alegran por mi. Se os quiere mucho.

No sabría cómo agradecerles a mis padres todo lo que me han dado y me siguen dando. Son mis pilares, mi fuerza. Solo decirles que les quiero muchísimos. A mis her-

manos Jose y Angela, que en estos cuatro años han cambiado tanto sus vidas. Muchas gracias por estar siempre ahí, estoy muy orgullosa de vosotros y no os puedo querer más de lo que os quiero.

A David, que más que nadie ha vivido conmigo este proceso, que sus ánimos y sus fuerzas han sido y son incalculables. Es imposible agradecer con palabras todo lo que me das, pero ten bien seguro que tienes todo mi corazón.

Summary

Introduction

The increasing availability of social data sources from socio-technological systems — systems that record our daily activity as credit card records, doctor databases, phone call records, email, etc.— and on-line social networks —as Facebook, Twitter, Instagram, etc.—, holds the promise to help us to understand social behavior from different perspectives. The study of all this data would give us a better knowledge about society and individuals; it could be used to adapt technology to social needs, improve services, improve transparency in political and social actions, move towards more cooperation and participation; but also it could lead to more corporate and governmental control and privacy problems. To address these topics we need to know how people interact among them: if they develop stable patterns or strategies in their actions, and how predictable these actions are. Note that if the data recorded is chaotic and random, we could never build anything on them. The aim of this thesis is precisely to uncover both structural and temporal patterns in social systems and to develop predictive models based on them.

Long-term evolution of statistical patterns

We investigate whether, despite the intricacies and randomness of the tie formation and decay processes at the microscopic level, there are macroscopic statistical regularities in the long-term evolution of social communication networks. Statistical regularities have indeed been reported in the long-term evolution of human organizations (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999) and human infrastructures such as the air transportation system (Gautreau et al. 2009); also in the activity patterns of single individuals, and are likely driven by daily and weekly periodicities (e.g. in communication (Barabási 2005; Oliveira and Barabási 2005; Malmgren et al. 2009a; Malmgren et al. 2010) and mobility (Brockmann et al. 2006; González et al. 2008)). However, due to the difficulty of tracking social interactions of a large pool of individuals for a long time, we still lack a clear picture of what statistical regularities emerge in the long-term evolution of social networks. In

particular, beyond relatively short periods of time of 12 to 18 months (Kossinets and Watts 2006; Miritello et al. 2013; Saramäki et al. 2014; Saramäki and Moro 2015), we do not know up to what extent social networks remain stable, or whether individuals change their social behavior with time.

To elucidate these questions, we analyze the evolution of an email network (Guimerà et al. 2003) of hundreds of individuals within an organization over a period of four consecutive years. To that aim, we characterize the long-term evolution of email communication networks in terms of the logarithmic growth rates (LGRs) on yearly bases of the weights of connections between users and the strengths of the users, where the logarithmic growth rate is a measure of the variation of the variables. We find that the LGRs distributions follow well-defined exponential decays for both the weights and strengths, and moreover, these distributions are stationary. These findings imply that fluctuations in connection weights and strengths are considerably larger than one would expect from a process with Gaussian-like fluctuations.

Next, we seek to better understand the evolution of the communication behaviour of individual users. Recent results suggest that the way individuals divide their communication effort among their contacts (their so-called “social signature”) is stable over the period of a few months (Saramäki et al. 2014). This is consistent with the hypothesis that humans have a limited capacity to simultaneously maintain a large number of social interactions (Dunbar 1998). Thus, the users develop different communication strategies of communication as it has been shown that some individuals tend to change their contacts frequently (“explorers”), whereas others tend to maintain contacts (“keepers”) (Miritello et al. 2013).

We investigate whether these differences exist at the scale of years and if individual communication strategies are stable in the long-term. To do so, we investigate three different aspects of the email communication: the standardized Shannon entropy of individuals communication, the turnover of contacts and the fraction of emails sent from a users to old contacts. We found that individuals have long-lasting social signatures and communication strategies.

Predicting collective and individual social behaviour

In this section we explore the possibility to predict behaviors based on these patterns. First, we analyze the predictability of the logarithmic growth rates (LGRs) for both weights and strengths based on the correlations with several network features and using well-performing machine learning algorithms.

To identify variables that correlate significantly with the variable to predict is important for dimension reduction and for prediction. However, finding variables significantly correlated with the variable to predict does not necessarily lead to predictive power (Lo et al. 2015). For the logarithmic growth rate, we find that besides strong correlation between the networks features, the LGRs is highly unpredictable. Moreover, we find that a simple linear approach considering the most correlated features

in the analysis, performs significantly better than a well-performing machine learning algorithm such as the Random Forest.

At the same time, we investigate up to what extent the social signature is associated with a specific user. To that end, we test if it is possible to reidentify users based on the communication features defined in the last chapter: the standardized Shannon entropy and the individual strength. We found that the standardized Shannon entropy and strengths could be used to distinguish users among them— with maximum average recall of 0.78 for the combination of the information from both the Shannon entropy and the strength. Also of importance is that the combination of different sources of information improve significantly the performance in terms of the average recall even when one of them is nearly uninformative.

Accurate and scalable social recommendation using mixed membership stochastic block models

With ever-increasing amounts of information available through online platforms, modeling and predicting individual preferences (e.g. on movies, books, songs and items in general) is becoming of great importance. Good predictions enable us to improve the advice to users, and reflect a better understanding of the socio-psychological processes that determine those preferences. We have developed a network-based recommender system that makes scalable and accurate predictions of individuals' preferences. Our approach is based on the assumption that there are groups of individuals and of items (movies, books, etc.), and that the preferences of individuals for items are determined by their group memberships, then the resulting overlapping communities are easily interpretable and meaningful. Importantly, we allow each individual and each item to belong simultaneously to different groups. The resulting overlapping communities and the predicted preferences can be inferred with a scalable maximum likelihood algorithm based on a variational approximation.

Our approach enables us to predict individual preferences in very large datasets with tens of millions of observed ratings. And regarding the predictability, our mixed-membership model is considerably more accurate than current algorithms available for such large datasets. In addition, using available demographic data of the users and the inferred parameters of the model, we are able to detect unveil trends on the users ratings profiles. For instance, we find that when the users are younger, their profiles are more homogeneous or similar among them than when they are older. Also, we perform this study for the different genders, finding that even though men are more similar among them when younger than in elder ages, this trend is more pronounced for women.

Temporal inference using mixed-membership tensorial stochastic block models

Most real networks are in constant evolution, and the increasing availability of time-resolved network data sources, e.g., from socio-technical systems and on-line social

networks, has brought to the forefront the need to study and understand time varying networks. Beside the relevance of the topic, the interest on the temporal inference is fairly recent and therefore there is little work in developing rigorous approaches. One of such approaches used Non-negative tensor factorization that uncovered reasonable group formations and time patterns on the scholar schedule (Gauvin L 2014). Also in (Schein et al. 2015) they predict international events using Bayesian Poisson Tensor Factorization. Another approach use Bayesian inference on layered stochastic block models (Peixoto 2015), in which he is able to find the hidden mesoscopic structure on time-varying networks by identifying the most meaningful time-binning to represent the networks.

Here we propose a new inference model, that is a tensorial Mixed-Membership Stochastic Block Model. The model assumes that users could belong simultaneously to different groups, but also the time-intervals could belong to different groups, where the interaction between groups of users and time-intervals is dominated by a block model like tensor. We used email data in a period of two months and with 65 users to validate the approach. Hiding the 20% of the data (events and non-events) the algorithm is able to distinguish accurately between events and non-events (with an AUC score of 0.88).

Conclusions

The work done in this thesis sheds light on the long-term stability of statistical regularities and patterns in communications networks, and on the predictability social networks. At the same time, we propose new models for prediction and inference that we validate with real data. The following conclusions can be drawn from the work:

- We have found that the long-term macro-evolution of email networks follows well-defined distributions, characterized by exponentially decaying log-variations of the weight of social ties and of individuals' social strength. These findings imply that fluctuations in connection weights and strengths are considerably larger than one would expect from a process with Gaussian-like fluctuations. Remarkably, together with these statistical regularities, we also observe that individuals have long-lasting social signatures and communication strategies.
- Our results suggest that the existence of correlations is not enough to build a satisfactory predictive model for the logarithmic growth rates. Regarding predictability, we found that a black box method such as Random Forest does not perform better than using the average expected growth for all the predictions.
On the other hand, we found that the individual standardized Shannon entropy and strengths could be used to reidentify users among them. Remarkably, we also found that the combination of different sources of information improves significantly the performance in terms of average recall; even when one of them is nearly meaningless.
- Our recommender model makes scalable predictions and is considerably more accurate than current algorithms for large datasets. Also, as the model is inter-

pretable, we have found that the parameters that maximize the likelihood allow us to infer trends for users and items given outside information. For the users we found interesting trends in age and gender: i) the younger are the users the more similar rating profiles they have between them; ii) the similarities in the ratings profiles for women among them and men among them follow this general trend in age; iii) the trend of the similarity with the age is much more pronounced for women than for men.

- We have developed a new algorithm for temporal inference, the mixed-membership tensorial stochastic block model, with a good performance in detecting real events. The model assumes mixed group membership on the users that interact among them and also on the time-intervals, where the interaction between groups is dominated by a block model tensor. We validate the model with email data, by performing AUC experiments with each of the training/test from the 5-fold cross-validation (hiding 20% of events and non-events). We get an average 0.88 AUC score which proves its classification power, even though the data is very sparse.

Contents

1	Introduction	1
1.1	Social networks	2
1.2	Communication networks	2
1.2.1	From local to global scale	3
1.2.2	From static to dynamic networks	4
1.3	Modeling recommendation systems	5
1.3.1	Inference methods	6
1.4	Scope of the work	10
2	Long-term evolution of statistical patterns	13
2.1	Introduction	13
2.2	Email data	14
2.3	The long-term evolution of email communication follows well-defined statistical patterns	16
2.3.1	Model selection and stationarity	17
2.3.2	Evolution of model parameters with time	18
2.4	Social signatures are stable in the long term	19
2.4.1	Analysis of how individuals distribute their communication	21

2.4.2	Analysis of the contacts turnover of users	25
2.4.3	Analysis of the fraction of emails to pre-existing contacts . . .	26
2.5	Discussion	28
3	Predicting collective and individual social behavior	29
3.1	Introduction	29
3.2	Logarithmic growth rates are largely unpredictable despite significant correlations	30
3.2.1	Predicting weights' LGRs $r_\omega(\mathbf{t} + 1)$	30
3.2.2	Predicting the strengts' LGRs $r_s(\mathbf{t} + 1)$	31
3.2.3	Leave-one-out experiments	33
3.3	Reidentifying users based on their social signature	35
3.3.1	Scores definition	36
3.3.2	Ranking and recall	37
3.4	Discussion	38
4	Social recommendation using mixed-membership stochastic block models	41
4.1	Introduction	41
4.2	Modeling ratings with a mixed-membership stochastic block model .	42
4.3	Benchmark algorithms	46
4.4	Ratings Data	48
4.5	Results	49
4.5.1	The MMSBM approach outperforms existing approaches . . .	49
4.5.2	The MMSBM approach provides a principled method to deal with the cold start problem	51
4.5.3	The MMSBM approach highlights the limitations of matrix factorization	52
4.6	Groups inferred with the MMSBM reflect trends of users and items . .	57
4.6.1	Inferred trends for movies	57
4.7	Discussion	60
5	Temporal inference using mixed-membership tensorial stochastic block models	63
5.1	Introduction	63
5.2	Modeling temporal inference with a mixed-membership tensorial stochastic block model	64
5.3	Results	67
5.3.1	The MMTSBM approach makes good predictions on hidden events	67
5.3.2	Groups inferred with the MMTSBM reflect temporal regularities	68
5.4	Discussion	70

6	Conclusions and perspectives	73
6.1	Conclusions	73
6.2	Perspectives	74
	References	77

1

Introduction

Social sciences as we know them nowadays, appear during the industrial revolution, where the need of labor in the factories moved people from the rural areas to the cities; as a result cities took a central role in societies which triggered the need to study and understand the changes in the social structure. The traditional focuses of social sciences include social stratification, social class, social mobility, religion, secularization, sexuality or deviance among others. Nowadays, the increasing availability of social data sources from socio-technical systems —systems that record our daily activity such as credit card records, doctor databases, phone call records, emails, etc.— and on-line social networks —such as Facebook, Twitter, Instagram, etc.—, holds the promise to help us to understand society from different perspectives. This so called 'data revolution' has opened a window for the study of social systems in an unprecedented manner.

Importantly, the understanding of all this data would give us a better knowledge about society and individuals; it could be used to adapt technology to social needs, improve services, improve transparency in political and social actions, move towards more cooperation and participation; but also it could lead to more corporate and governmental control and privacy problems. To address these topics we need to know how people interact among them: if they develop stable patterns or strategies in their actions, and how predictable these actions are. Note that if the data recorded is chaotic and random, we could never build anything on them.

1.1 Social networks

Any social system is driven by people (or groups of people) and their interactions. The complete knowledge of the individual people do not explain the emergent properties of the system (such as the evolution, functionalities, social roles, etc.), given that the interactions between individuals carry important information. The network approach is a simplification/representation of the social system by nodes (people) and connections (interactions) that has successfully capture some of its complexities. The study of networks if social interactions is in fact a field with a long standing tradition in the social sciences.

The application of networks to social analysis started in the 1930s, but it was not until the 1980s that the social networks became settled in the social and behavioral sciences. On the origins, mainly by sociologist, anthropologist and psychologist, social networks were focused on identifying communities and structures of small groups (Lévi-Strauss 1947; Barnes 1954; Nadel 1957); then the mathematical formulation of networks extends the approach to study the properties of these structures (Blau 1960; Wellman 1988); and more lately the longitudinal studies on with larger datasets establish the bases for the development of small world theory (Stanley Milgram developed the "six degrees of separation" thesis), scale-free networks theory and dynamical models (S. H. Strogatz 1998; Barabási and Bonabeau 2003; Ebel H. and S. 2002).

Interestingly, the availability of data from a variety of sources enables us to build networks that reflect the social behaviour in different contexts. For instance, social connections networks, such as Facebook or Twitter, where connections indicate mostly friendship, but where the posts are broadcast to all the friends at the same time; multi-media sharing such as Youtube or Flirck, where piece of information are freely broadcast; dyadic communication networks such as phone call network or emails, where the information is from one user to an other at a time; or recommender systems such as Netflix, MovieLens or Amazon reviews, where the users post their preferences on different items –represented as bipartite network with two types of nodes, users and items—. Particularly, in this thesis we will deal with communication networks and recommender systems from different perspectives.

1.2 Communication networks

Communication networks are defined as those networks where the connections between nodes represent the exchange of information between pairs of users at a time. Common examples of communication networks are phone call networks or email networks. The study of these networks has shed light on relevant problems as epidemic spreading, information flows or online-security among others (Liljeros et al. 2001; Liljeros et al. 2003; Balcan et al. 2009; Kossinets et al. 2008; Colizza et al. 2006). Communication networks are time resolved, meaning that the social events are recorded in time. This fact allows us to analyze communication networks from different perspectives. We can study the network as a whole, looking at its communication flows, global evolution, structure of the network (Kossinets et al. 2008; Gautreau et al. 2009;

Jackson and Watts 2002; Barabási et al. 2002), and also focus on the individuals communication as for example, studying individual communication strategies, links formation, modeling individuals communication (Saramäki et al. 2014; Kossinets and Watts 2006; Jo et al. 2012; Malmgren et al. 2008). On the other hand, we can potentially study time resolved networks as static networks, aggregating the temporal information to study concepts such as small-world effect, degree distributions, clustering, random graph models, or preferential attachment (Newman 2003; Newman 2001; Guimerà et al. 2005), or as dynamic networks, taking into account the temporal dimension of the events to study temporal social patterns, long-term evolution, growth of the networks or temporal inference among others (Iribarren and Moro 2009; Miritello et al. 2013; Nowak 2006; Gauvin L 2014).

1.2.1 From local to global scale

Most of the time, the complete knowledge of the individual behaviors do not explain the global trends or patterns of the communication networks, or the other way around, it is impossible to infer the individual communication strategies from the global trends.

The analysis of global trends in social networks has been driven for different interests such as trades among nations, international political relations, or more recently world wide social networks as Facebook or Twitter (Gautreau et al. 2009; Schein et al. 2015). The basic conclusion from the research on this field until now is that, beside the apparent randomness of social behavior at the microscale, at the organizational level different studies show that macroscopically there are collective patterns that could be characterized and modeled. For instance it has been proved the universality of different interpersonal communication processes such as the letter correspondence and the email (Malmgren et al. 2009b), where the distributions of inter-event times for two different networks are similar after rescaling. Also at the organizational level, it has been shown that very different human activities present similar exponential decays in their growth as companies sales or air transportation networks (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999; Gautreau et al. 2009). As an illustration, consider the air transportation network, which is a weighted network, where the connections' weights are defined as the number of flights between i and j airports w_{ij} . Then, the logarithmic growth rate of the weights is defined as $r_{ij}(t) = \log(w_{ij}(t+1)/w_{ij}(t))$. Fig. 1.1 shows that the distributions of growth displays an exponential decay, implying that fluctuations in connection weights are considerably larger than one would expect from a process with Gaussian-like fluctuations.

Statistical regularities have also been reported in the activity patterns at the single individuals level. It has been shown that individual dyadic communication networks is described by the circadian cycle/task repetition (Malmgren et al. 2009) and, at the same time, characterized by bursts of rapidly occurring events separated by long periods of inactivity (Barabási 2005). In fact, the process of individual e-mail communication has been modeled as a cascading non-homogeneous Poisson process that takes into account both periodic and the bursty nature of communication (Malmgren et al. 2008).

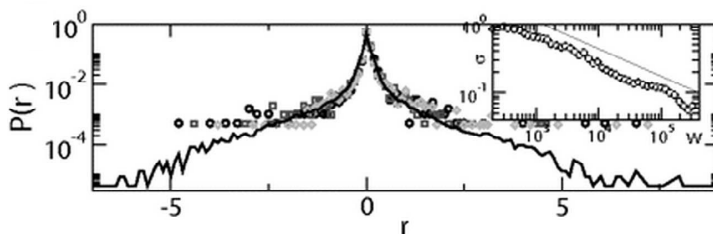


Figure 1.1: **Logarithmic growth rate distributions of the airports traffic.** Distribution of the monthly growth rates $r = \log(w(t+1)/w(t))$. The full line corresponds to the distribution obtained over the 11 years under study. Symbols correspond to 3 single months. The Inset shows the standard deviation σ of the conditional distributions $P(r|w)$ as a function of w , showing that the distribution of the growth rates become narrower for larger weights. Original source: (Gautreau et al. 2009). *Reprinted with permission.*

Also, there is a recent interest in identifying patterns of the individual communication activity. The existence of these patterns is based on the hypothesis that social communication is constrained in some way either by cognition or by the time individuals have available for social interaction, or both; then individuals appear to have a social signature in the way they communicate with the rest of the network (Saramäki et al. 2014). Based on some personal survey and phone call records, they found that users allocate more attention in some contacts than others depending on the strength of the personal link among them. Although social signatures vary between individuals, a given individual appears to retain a specific social signature over time. In this thesis we explore the possibility that social signature is also present in the email communication, which may be an indicator of universal mechanisms underlying these different communication processes.

1.2.2 From static to dynamic networks

Real social networks are dynamic networks, where social agents interact within time. Until recently, most of the studies in networks have been on static networks (mainly because of the difficulty of recording datasets that prolong enough in time) (Albert et al. 1999; Barrat and Barthélemy ; J-P Onnela 2007). Dynamic studies are more into communication spread and evolution of the network, while static networks are more common for the study of network structures. When studying the structure (cohesion, modularity, group membership or communities) it is common to aggregate the network, assuming that the structure is stable, and also over estimating correlations between temporal events, or in other words, assuming complete correlations (Girvan and Newman 2002; Newman 2004).

The assumption that the network is not changing in time is plausible and needed when predicting, since if the network is changing past data would not give any clue of the future behavior. As an example, in recommender systems while predicting ratings

from users to items, the algorithms make good predictions under the assumption that the preferences of users has not change in time (Sarwar et al. 2001; Guimerà et al. 2012; Koren et al. 2009). Of course it will not be valid forever, it would depend on the time scale of the problem to study. For instance, one may want to aggregate the data in small time intervals, but large enough to perform statistical and unveil temporal patterns (if there are), then look at a larger time scale to analyze the stability of these patterns. The recent existence of massive longitudinal dataset allows us to assess for the first time stability and long-term evolution analysis on networks (Godoy-Lorite et al. 2015). The assumption that the events are completely correlated between them could also be misleading. Looking at the communication channels, Gueorgi Kossinets et al. (Kossinets et al. 2008) found that *network backbone* resulting from the temporal network—that are channels where the information has the potential to flow quickest— are different from the structure resulting from the static network. As a conclusion, it is increasingly clear that to study the system as a static or dynamic system should be adapted to the particular problem to address.

1.3 Modeling recommendation systems

A widely studied example of individual behavior is the case of recommender systems. In these systems, users rate items within a limited scale of possible ratings, said ratings from 1 to 5 integers. The prediction is simpler than in other systems, since the number of solutions is limited. The two main approaches of the problem are collaborative filtering (CF) and content-based filtering. Content-based algorithms use external information about users or items to predict. This external information could be demographic features for the users, past habits of users, finite categories that describe the items or top selling items; however, CF is the most successful recommendation technique to date. Many CF algorithms have been used in measuring user similarity or item similarity in recommender systems (Sarwar et al. 2001; Resnick and Riedl 1994). Another very popular recommender algorithm is matrix factorization (MF), which consists in decomposing the ratings matrix (the matrix of the actual ratings of users to items) into two different matrices, one for the users that represent the user 'features', and the other for the items which also represent their 'features' values (Koren et al. 2009; Paterek 2007; Funk 2006). MF performs a very good accuracy in the predictions an also is very efficient computationally. However, until now and as far as we know, the model that performs better accuracy in the prediction is the Bayesian stochastic block model (SBM) (Guimerà et al. 2012). The SBM is a recommender algorithm based on the generative model for random graph of the same name, that will also be relevant for other models developed in this thesis.

Stochastic block models

The Stochastic block model, as said, is a family of generative models M_{BM} , where the nodes are partitioned into groups α, β, \dots , and where the probability that two nodes are connected depends only in the groups to which they belong $Q_{\alpha\beta} \in [0, 1]$ (see Fig. 1.2).

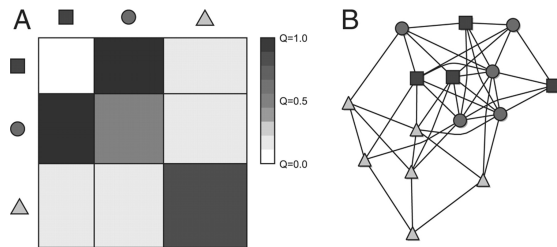


Figure 1.2: **Stochastic block models.** A stochastic block model is fully specified by a partition of nodes into groups and a matrix Q in which each element $Q_{\alpha\beta}$ represents the probability that a node in group α connects to a node in group β . (A), A simple matrix of probabilities Q . Nodes are divided in three groups (which contain 4, 5, and 6 nodes, respectively) and are represented as squares, circles, and triangles depending on their group. The value of each element $Q_{\alpha\beta}$ is indicated by the shade of gray. (B) A realization of the model in A. Original source: (Guimerà and Sales-Pardo 2009). *Reprinted with permission.*

Note that this model is suitable to describe directed, undirected and bipartite networks and is flexible enough to generate any possible network structure. As for example: if the probability matrix is constant $Q_{\alpha\beta} = p$, then the result is the Erdős–Rényi model; if $Q_{\alpha\alpha} > Q_{\beta\delta}$, $\beta \neq \delta$, then the network is assortative; and if $Q_{\alpha\alpha} < Q_{\beta\delta}$, $\beta \neq \delta$ then the network is disassortative.

In the case of a recommender systems we have a bipartite network with two types of nodes (users and items) where the nodes of one type could only interact with nodes of the other type. Thus, the block model matrix $Q_{\alpha\beta}$ is also a two dimensional matrix, but now α is the users' group and β the items' group. The ratings are treated as independent labels for the connections between users and items, as a consequence there are as many independent block model matrices as possible rating in the system $Q_{\alpha\beta}^r$, representing the probability of a node in group α to rate r to an items of group β . The SBM recommender algorithm uses Monte Carlo sampling over the most relevant block models that represents the real ratings system (Guimerà et al. 2012).

1.3.1 Inference methods

Within the recommendation algorithms, the method for model selection is of central importance. Some very good models could become useless if they are computationally unfeasible, and the other way around, a model with modest results could become very popular if it is computationally efficient. Obviously, the objective is to get the best possible accuracy using the less computational effort. Parameter free models are not necessarily faster algorithms, since they could require a large number of operations to make the predictions. For the parametric and nonparametric models —The difference between parametric models and non-parametric models is that the former has a fixed number of parameters, while the latter grows the number of parameters with the amount of training data—, the computational time depends on the difficulty arriving

to the model parameters (or set of parameters) that computes the best possible predictions. There are several approaches to do so, depending on the particularities of each model. The general problem to solve is given a dataset of observables D , we want to find a model $m(\theta)$, where θ are the parameters of the model, that predict x unobservables. Then the probability of the prediction as a function of the observables and the model is,

$$P(x|D, m(\theta)) = \int d\theta P(x|\theta, D)P(\theta|D, m). \quad (1.1)$$

where $P(\theta|D, m)$ is the *a posteriori* probability (or posterior). Applying Bayes theorem we have that,

$$P(\theta|D, m) = \frac{P(D|\theta, m)P(\theta, m)}{P(D)} \quad (1.2)$$

Where $P(D|\theta, m)$ is the likelihood of the model, and $P(\theta, m)$ is the *a priori* (or prior) probability that the model is the correct one before collecting the observation data.

Our goal is to solve the prediction problem from Eq. 1.1, also called the predictive distribution, but can be very high-dimensional and difficult to compute. Therefore, there are several methods to approximate its value.

Monte Carlo methods

Monte Carlo methods approximate the integral in Eq. 1.1 by random sampling on the model parameters. When the probability distribution of the variable is too complex, a popular option is to use a Markov Chain Monte Carlo (MCMC) sampler. The central idea of MCMC approach is to design a judicious Markov chain model with a prescribed stationary probability distribution. A Markov process is a random process for which the future (the next state of the model parameters) depends only on the present state. It has no memory of how the present state was reached. The integral is approximated by the empirical measures of the random states of the MCMC sampler. This is ensured thanks by the ergodic theorem, which requires that every state must be aperiodic: the system does not return to the same state at fixed intervals, and be positive recurrent: the expected number of steps for returning to the same state is finite. There are different algorithms for the sampling: Metropolis–Hastings algorithm, Gibbs sampling, Wang and Landau algorithm, Sequential Monte Carlo samplers, etc. We will explain with some detail some of the most relevant of them.

The Gibbs sampling, generates a Markov chain of samples, each of which is correlated with nearby samples. Suppose we want to generate S samples of the system states $\theta^t = \{\theta_1^t, \dots, \theta_N^t\}$ (or parameters values) from a joint distribution $P(\theta_1, \dots, \theta_N)$. To do so, we start the sampling from some initial values θ^0 . Then, to get the next sample (call it the $t + 1$ -th sample for generality) we sample each component variable $\theta_i^{(t+1)}$ from the distribution of that variable conditioned on all other variables, making use of the most recent values and updating the variable with its new value as soon as it has been sampled. This requires updating each of the component variables in turn. If we are up

to the i -th component we update it according to the distribution specified by,

$$P(\theta_i | \theta_1^{(t+1)}, \dots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \dots, \theta_N^{(t)}). \quad (1.3)$$

To get the whole sample, repeat the above step for all the parameters S times. As a result, care must be taken if independent samples are desired typically by thinning the resulting chain of samples by only taking every n -th values. Note that samples from the beginning of the chain (the burn-in period) may not accurately represent the desired distribution.

Metropolis-Hastings is a widely used MCMC sampling method. In this case, the sampling over the different states of the systems (or parameters values) for a general probability $P(\theta)$, is speared into two steps: the proposal of a new state and the acceptance-rejection of that proposal. The proposal distribution $g(\theta^t \rightarrow \theta^{t+1})$ is the conditional probability of proposing a state θ^{t+1} given θ^t . The proposal may be some random change, for instance in the stochastic block model recommender it is the random change of a user from one group to another. The acceptance distribution $A(\theta^t \rightarrow \theta^{t+1})$ is the conditional probability to accept the proposed state θ^{t+1} , which 'controls' if the proposed state is more likely than the previous one. Assuming detailed balance—that is a sufficient but not necessary conditions, which implies that the probability of being in state θ^t and transitioning to state θ^{t+1} must be equal to the probability of being in state θ^{t+1} and transitioning to state θ^t —, the new proposed state θ^{t+1} is accepted with probability,

$$A(\theta^t \rightarrow \theta^{t+1}) = \min \left(\frac{P(\theta^{t+1})g(\theta^{t+1} \rightarrow \theta^t)}{P(\theta^t)g(\theta^t \rightarrow \theta^{t+1})} \right). \quad (1.4)$$

As well as in the Gibbs sampling, in order to ensure independent samples, we only consider the n -th accepted state to the sampling. The acceptance step ensures that the most relevant states to contribute to the sampling.

Variational approach

The key to the variational method is to approximate the integral in Eq. 1.1 with a simpler form that is tractable, forming a lower or upper bound through a distribution function $q(\theta)$. This is done using of the Jensen's inequality to the logarithm of the predictive distribution Eq 1.1—the logarithm is a monotonous increasing function, then to maximize a function is equivalent that to maximize the logarithm of this function, which is convenient for this case.

$$\begin{aligned} \log P(x|D, m) &= \log \int d\theta q(\theta) \frac{P(x|\theta, D)P(D|\theta, m)P(\theta|m)}{q(\theta)} \\ &\geq \int d\theta q(\theta) \log \frac{P(x|\theta, D)P(D|\theta, m)P(\theta|m)}{q(\theta)} \equiv F_x(q, \theta). \end{aligned} \quad (1.5)$$

The variational problem aims at finding the functional form of $q(\theta)$ that maximize the integrand, which seems to be costly since in general it implies a sampling over

the space of distribution functions. But it could be easily shown that the function that holds the equality in Eq. 1.5 is of the form of $q(\theta) \propto P(x|\theta, D)P(D|\theta, m)P(\theta|m)$, divided by some normalization factor, since it is a distribution function. The standard procedure to find the $q(\theta)$ that maximizes $F_x(q, \theta)$, is to derive $F_x(q, \theta)$ as a function of $q(\theta)$ and make it equal to zero. There are several variational methods and theorems to solve Eq. 1.5 integral, such as mean-field approximation, variational Bayesian Expectation-Maximization (in case there are hidden variables), conjugate-exponential method, Cheeseman-Stutz approximation, also methods to optimize the priors, etcetera.

One of the most common (and simplest) approaches assumes that the priors are Dirac delta functions $P(\theta|m) = \delta(\theta - \theta^*)$, where θ^* are the values of the parameters in the maximum likelihood, thus,

$$\log P(x|D, m) = \log \int d\theta P(x|\theta, D)P(\theta|D, m) \simeq \log P(x|\theta^*, D)P(\theta^*|D, m). \quad (1.6)$$

In this case the prediction problem becomes a maximum likelihood (ML) problem. The variational approach for the ML problem would be,

$$\begin{aligned} \mathcal{L} &= \log P(D|m) = \log \int d\theta P(D|\theta, m)\delta(\theta - \theta^*) \\ &= \log q(\theta^*) \frac{P(D|\theta^*, m)}{q(\theta^*)} \geq q(\theta^*) \log \frac{P(D|\theta^*, m)}{q(\theta^*)} \equiv F(q, \theta^*), \end{aligned} \quad (1.7)$$

where $q(\theta)$ is the variational distribution function for the maximum likelihood problem; from all possible distribution functions, we want to find $q^*(\theta)$ that maximized the likelihood in Eq. 1.7. To do so, we derive $F(q, \theta)$ as a function of q with fixed parameters θ and make it equal zero. That results in

$$q^*(\theta) = \frac{P(D|\theta, m)}{N}, \quad (1.8)$$

where N is a normalization factor that ensures that $q(\theta)$ is a distribution function. Once we know $q^*(\theta)$, we want to find the parameters θ^* that maximize $F(q, \theta)$. To that aim we derive $F(q, \theta)$ as a function of the parameters and make it equal zero,

$$\frac{\partial F(q, \theta)}{\partial \theta} = 0, \quad \forall \theta. \quad (1.9)$$

From the solution of these equations (one for each parameter of the system $\theta_i^* = f(q^*, \theta_{j \neq i}^*)$), we get a system of update equations. Therefore, the values for the parameters in the maximum-likelihood are found through the following iterative process: 1) initialize randomly the parameters θ^t ; 2) compute q^t as a function of θ^t using Eq. 1.8; 3) compute the new values of the parameters θ_i^{t+1} as a function of q^t and the other parameters values in the last step $\theta_{j \neq i}^t$ using the update equations 1.9. Alternatively, one can: 1) initialize randomly q^t ; 2) compute θ^t using the update equations 1.9; 3)

compute q^{t+1} as a function of θ^t using Eq 1.8. In both cases, it is necessary to iterate 2 and 3 until convergence. At each step it is ensured that $F(q, \theta)$ grows.

If there are constraints over the parameters, for instance $\int d\theta \theta = 1$, one can use the Lagrange multipliers method. Then the function to maximize becomes,

$$\tilde{F}(q(\theta), \theta) = F(q, \theta) - (\lambda(\int d\theta (\theta - 1))), \forall \theta. \quad (1.10)$$

Importantly, the same approach could be used for parametric problems, where the number of parameters is fixed, as we will see in this thesis. Note that, when converting a predictive problem (the same for the maximum a posterior problem) into a ML problem, even each interaction improves the true likelihood, the improvement is only in the direction of a local maximum. For this reason, a typical approach is to make a sampling with different initial conditions to explore the sampling parameters space. If the likelihood is well-behaved, the sampling over different initial conditions would be enough to get a good result, but it could become more complicated when increasing the dimension of the problem. In contrast to Metropolis-Hastings, that compute the sampling over all relevant states, the ML solution from one initial conditions is just one local maximum. On the other hand, while Metropolis sampling is costly computationally, the variational maximum likelihood (VML) approach is computationally more efficient since it finds the solution only in the direction of the maximum likelihood. There are other methods for finding maximum likelihood estimates, such as gradient descent, conjugate gradient or variations of the Gauss–Newton method. Unlike VML, such methods typically require the evaluation of first and/or second derivatives of the likelihood function.

1.4 Scope of the work

The main interest in this thesis has been to get a deeper understanding on the statistical patterns and predictability of social networks. This thesis combine the statistical analysis with the development of new models and the validation with real data.

We start analyzing the evolution of an email network conducted on longitudinal email data of hundreds of individuals within an organization over a period of four consecutive years. Besides statistical regularity in email networks has been reported, still it is unknown if these regularities remain stables in the long-term, given the difficulty of having social datasets that prolong enough in time (there are not many dataset beyond relatively short periods of time of more than 18 months). In the analysis of email data, we find statistical regularities in the long-term at a global and local scale (Chapter 2).

Given the statistical regularities found in Chapter 2, we explore the possibility to predict the global exchange of emails based on these patterns. To do so, we analyze the correlations of the email growth with different network features. From these correlations, we make estimations on the future growth using, among others, well-performing machine learning algorithms. At the same time, we also analyze until which point the individuals' communication strategies could make users indistinguishable. To that aim, we try to reidentify users only based on their email correspondence (Chapter 3).

Next, we develop a network-based recommender system that makes scalable and accurate predictions of individuals' preferences (Chapter 4). To make good recommendations is of increasing importance nowadays with the growth of options available through online platforms. Our approach is based on the assumption that there are groups of individuals and of items (items, books, etc.), and that the preferences of an individual for an item are determined by their group memberships. Importantly, we allow each individual and each item to belong simultaneously to different groups. The resulting overlapping communities can be inferred with a scalable algorithm based on a variational approximation, which enables us to make accurate predictions on datasets with tens of millions of ratings. Moreover, our algorithm outperform current recommender algorithms.

Finally, we develop a mathematical model for the inference of events in time (Chapter 5). Most of the real networks are temporal networks, and beside the importance in the understanding of inner dynamics that results in the observations of the systems, the interest in this topic is very recent—mostly due to the lack of temporal longitudinal data. In our time inference model, the users also belong to different groups, but importantly, the algorithm also perform group membership on the time intervals. We perform a cross validation of the model using email data, where the algorithm results in distinguish accurately real events from no-event.

Concluding remarks and perspectives for future work are given in the last chapter of the thesis.

Long-term evolution of statistical patterns

2.1 Introduction

Individuals thrive in a social environment through the construction of social networks. Ties in these networks satisfy individual needs and are necessary for well-being, but the effort, time and cognitive investment that each tie requires limit the ability of individuals to maintain them (Miritello et al. 2013; Miritello et al. 2013; Saramäki et al. 2014). As a result of this limit, social networks are intrinsically dynamical, with individuals constantly dropping ties and replacing them by new ones (Miritello et al. 2013; Miritello et al. 2013; Kossinets and Watts 2006).

Several factors are known to play an important role in the intricate microscopic process of tie replacement—for example, mechanisms such as homophily (McPherson et al. 2001) and triadic closure (Easley and Kleinberg 2010) have been found to generally drive tie creation (Kossinets and Watts 2006). However, these processes are remarkably noisy (Kossinets and Watts 2006) and are modulated by the distinct social behaviors of each individual (Miritello et al. 2013; Miritello et al. 2013; Saramäki et al. 2014), so that in the short term individual ties appear and decay in a highly unpredictable fashion.

In this chapter we investigate whether, despite the intricacies and randomness of the tie formation and decay processes at the microscopic level, there are macroscopic statistical regularities in the long-term evolution of social communication networks. Statistical regularities have indeed been reported in the activity patterns of single individuals, and are likely driven by daily and weekly periodicities (e.g. in communication (Barabási 2005; Oliveira and Barabási 2005; Malmgren et al. 2009a; Malmgren

et al. 2010) and mobility (Brockmann et al. 2006; González et al. 2008)); statistical regularities have also been reported in the long-term evolution of human organizations (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999) and human infrastructures such as the air transportation system (Gautreau et al. 2009). However, due to the difficulty of tracking social interactions of a large pool of individuals for a long time, we still lack a clear picture of what statistical regularities emerge in the long-term evolution of social networks. In particular, beyond relatively short periods of time of 12 to 18 months (Kossinets and Watts 2006; Miritello et al. 2013; Saramäki et al. 2014; Saramäki and Moro 2015), we do not know up to what extent social networks remain stable, or whether individuals change their social behavior with time.

Besides the academic interest of these questions, they are also of practical relevance because the structure of social networks plays an important role in processes such as the spread of information or epidemics (Liljeros et al. 2001; Liljeros et al. 2003; Balcan et al. 2009). The static analysis of communication networks has shed light on some important aspects (e.g. the role of weak ties in keeping the stability of social networks (J-P Onnela 2007)). However, it is increasingly clear that ignoring network dynamics can lead to very poor models of collective social behavior, and that even fluctuations at a microscopic level often have a large impact on social processes (Iribarren and Moro 2009).

To elucidate these questions, here we analyze the evolution of an email network (Guimerà et al. 2003) of hundreds of individuals within an organization over a period of four consecutive years. We find that the macro-evolution of social communication networks follows well-defined statistical patterns, characterized by exponentially decaying log-variations of the weight of social ties and of individuals' social strength. At the same time, we find that individuals have long-lasting social signatures and communication strategies.

2.2 Email data

We analyze the email network of a large organization with over 1,000 individuals for four consecutive years (2007-2010). For this period, we have information of the sender, the receiver and the time stamp of all the emails sent within the organization using the corporate email address. To preserve users' privacy, individuals are completely anonymized and we do not have access to email content. The email networks for each year comprise $n_{2007} = 1,081$, $n_{2008} = 1,240$, $n_{2009} = 1,386$, and $n_{2010} = 1,522$ individuals. The total number of emails recorded each year is $l_{2007} = 211,039$, $l_{2008} = 303,619$, $l_{2009} = 368,692$, and $l_{2010} = 444,493$.

Since the number of emails sent from i to j during a year is typically similar to the number of emails sent from j to i , we consider the undirected weighted network in which the weight ω_{ij} of the connection between users (i, j) represents the total number of emails exchanged by this pair of users during one year. In Figure 2.1 the

very high Pearson's correlation coefficient of $\rho = 0.83$, $p < 10^{-323}$ confirms the visually apparent linear relationship.

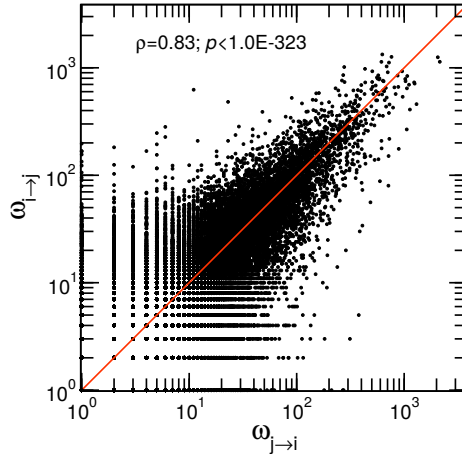


Figure 2.1: **Equivalence of the direct and undirected email network.** Scatter plot of the $\omega_{i \rightarrow j}$, the number of emails that user i sends to user j during one year versus $\omega_{j \rightarrow i}$. The coordinates of each pair have been randomly given in order to avoid any bias. The red line represents the situation in which all the emails sent were answered.

Because we are interested in non-spurious social relationships, in our analysis we only consider connections with weight $\omega_{ij} \geq 12$, that is we only consider connections between pairs of users that exchange at least an email per month on average. Such filters are known to generate networks whose connections resemble more closely self-reported social ties (Wuchty and Uzzi 2011).

Ethics statement

Our work is exempt from IRB review because: i) The research involves the study of existing data—email logs from 2007 to 2010, which the IT service of the organization archived routinely, as mandated by law; ii) The information is recorded by the investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. Indeed, subjects were assigned a "hash" by the IT service prior to the start of our research, so that none of the investigators can link the "hash" back to the subject. We have no demographic information of any kind, so de-anonymization is also impossible. Finally, we do not report results for any individual subject (or even for groups of users), but only aggregated results for all users.

2.3 The long-term evolution of email communication follows well-defined statistical patterns

We characterize the long-term evolution of email communication networks in terms of two properties: the weight $\omega_{ij}(t)$ of connections for year t (Fig. 2.2); and the user strength $s_i(t) = \sum_j \omega_{ij}(t)$ (Fig. 2.3) (Barrat and Barthélemy), that is, the total number of emails exchanged by each user i during year t .

The distributions of connection weights and user strengths have two remarkable features (Figs. 2.2A and 2.3A). First, these distributions are fat-tailed, with values spanning over three orders of magnitude. Second, these distributions are stable for the four years we study (despite a small but significant shift towards higher number of emails).

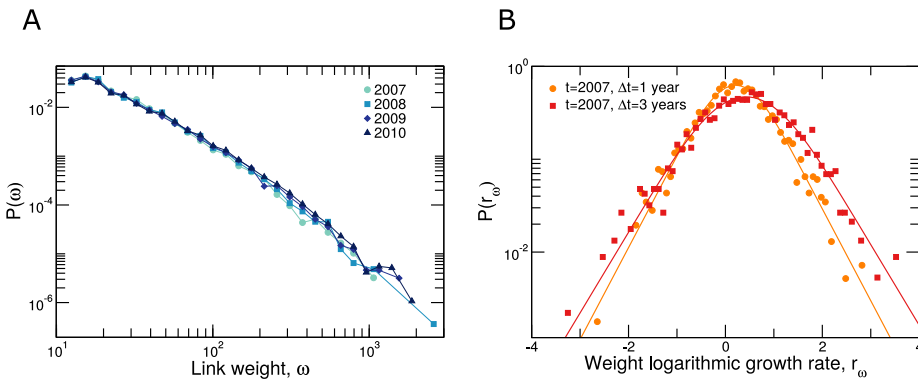


Figure 2.2: **Time evolution of connections' weights.** The weight ω_{ij} of a connection between users (i, j) corresponds to the number of emails exchanged by i and j during a whole year. We only consider connections with $\omega \geq 12$ (see text) (A) Distributions of weights for each one of the years in our dataset (2007-2010). Note that the distribution is stable in time. (B) Distribution of the weight logarithmic growth rates $r_\omega = \log(\omega(t + \Delta t)) - \log(\omega(t))$ for $t = 2007$ and $\Delta t = 1, 3$ (dots and squares, respectively). Lines show fits to the convolution of a Laplace distribution and a Gaussian distributed noise (see Eq. (2.6)). Note that as Δt increases the distributions are slightly wider, the peaks are rounder and shift to the right.

Besides the overall stability of the distributions, we observe a large variation in connection weights and user strengths from year to year. To characterize this variation, we define the logarithmic growth rates (LGRs) (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999; Gautreau et al. 2009)

$$r_\omega(t, \Delta t) = \log \left(\frac{\omega(t + \Delta t)}{\omega(t)} \right) \quad (2.1)$$

$$r_s(t, \Delta t) = \log \left(\frac{s(t + \Delta t)}{s(t)} \right), \quad (2.2)$$

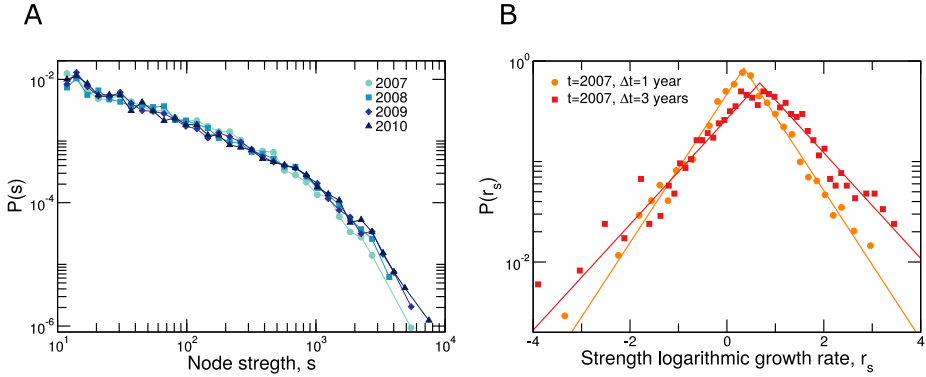


Figure 2.3: **Time evolution of nodes' strengths.** The strength s_i of node i is the number of emails that user i exchanged with other users during one year. **(A)** Distributions of strengths for each one of the years in our dataset (2007-2010). Note that the distribution is stable in time. **(B)** Distribution of strength logarithmic growth rates $r_s = \log(s(t + \Delta t)) - \log(s(t))$ for $\Delta t = 1, 3$ years (dots and squares, respectively). Lines show fits to a Laplace distribution (Eq. 2.3). Note that as Δt increases the distributions are slightly wider and the peaks shift to the right.

and study their distributions (Figs. 2.2B and 2.3B as examples for weights' and strengths' LGRs; LGRs for all $t = 2007, 2008, 2009, 2010$ and $\Delta t = 1, 2, 3$ years are addressed below). These distributions are tent-shaped and have exponentially decaying tails.

2.3.1 Model selection and stationarity

We analyze the evolution of email communication flows through the logarithmic growth rates of link weights and node strengths (LGRs). We consider the following models for the LGR distributions (Figs. 2.4 and 2.5):

- a Laplace or symmetric exponential distribution

$$P_L(r|\sigma_{\text{exp}}) = \frac{\exp(-|r - \mu|/\sigma_{\text{exp}})}{2\sigma_{\text{exp}}}; \quad (2.3)$$

- a Gaussian distribution

$$P_{\text{Gauss}}(r|\sigma_G) = \frac{e^{-(r-\mu)^2/2\sigma_G^2}}{\sigma_G\sqrt{2\pi}}; \quad (2.4)$$

- an asymmetric Laplace distribution

$$P_{\text{asymm-L}}(r|\sigma_{\text{left}}, \sigma_{\text{right}}) = \begin{cases} \frac{\exp(-|r-\mu|/\sigma_{\text{left}})}{\sigma_{\text{left}} + \sigma_{\text{right}}} & \text{if } r \leq 0 \\ \frac{\exp(-|r-\mu|/\sigma_{\text{right}})}{\sigma_{\text{left}} + \sigma_{\text{right}}} & \text{if } r > 0 \end{cases}; \quad (2.5)$$

- a convolution of a Laplace and a normal distribution

$$P_{conv}(r|\sigma_{\text{exp}}, \sigma_G) = \int_{-\infty}^{\infty} \frac{e^{-|\rho|/\sigma_{\text{exp}}}}{2\sigma_{\text{exp}}} \frac{e^{-(r-\mu-\rho)^2/2\sigma_G^2}}{\sigma_G\sqrt{2\pi}} d\rho. \quad (2.6)$$

We choose the Bayesian information criteria (BIC) to establish which is the best model for the distributions of LGRs (Schwarz 1978).

For user strengths, a Laplace distribution Eq. 2.3 provides the best overall fit to the data. For connection weights a pure Laplace distribution does not provide a good fit to the data. Since the majority of points are located around the rounding mode of distribution, while maximizing the likelihood for the Laplacian fit (or minimizing the BIC) it receives much of the weight from these points around the mode and fails on capture real exponential parameter of the tails. In order to solve this problem, we assume that the observed rate r is a combination $r = \tilde{r} + \epsilon$, where \tilde{r} is Laplace distributed according to Eq. (2.3) and ϵ is a normally distributed “noise”, so that $P(r_\omega)$ is the convolution of a Laplace and a normal distribution (Eq. 2.6). In this model, the Gaussian noise do not perturb the exponential tails because for large values of r these perturbations are negligible, but around the peak for small values of r the noise do affect, rounding the peaks of the distributions.

However, for fixed Δt , the mode $\mu(t, \Delta t)$ of the distribution changes slightly to the starting year $t = 2007, 2008, 2009$ (see Fig. 2.6), especially for $t = 2007$ and not significant for $t = 2008$ and $t = 2009$. Therefore in order to assess if the functional form of LGR distributions is stationary (that is with different modes but otherwise with the same model parameters), we need to compare the distribution of centered LGRs, $r^0 = r - \mu(t, \Delta t)$, for fixed Δt . Table 2.3.1 shows the results of comparing pairs of distributions using the Kolmogorov-Smirnov test. According to test results, at a 1% significance level, we cannot reject the hypothesis that for fixed Δt distributions of centered logarithmic growth rates of strengths for different years come from a single distribution. Note that for $\Delta t = 1$ a multiple testing correction to the p-value, would further strengthen our results.

Therefore in Fig. 2.7, for a fixed Δt the distributions of r_ω^0 and r_s^0 are pooled together in a single curve. Note also that the same functional form that describes growth rates from one year to the next, $\Delta t = 1$ year, also describes growth rates at $\Delta t = 2$ years and $\Delta t = 3$ years.

2.3.2 Evolution of model parameters with time

In Fig. 2.7 we show the distributions of the centered logarithmic growth rates for both weights and strengths. According to the Bayesian Information Criterion (BIC) the best fit is for $P(r_\omega^0)$ a convolution of a Laplace distribution and a Gaussian (Eq. 2.6) and for $P(r_s^0)$ a Laplace distribution (Eq. 2.3). We estimate the parameters using maximum likelihood for the best model in each case (Fig. 2.8).

In general we find that as Δt increases, the exponential tails becomes wider and the total density of exchanged emails increases for both weights and strengths. For $P(r_\omega^0)$

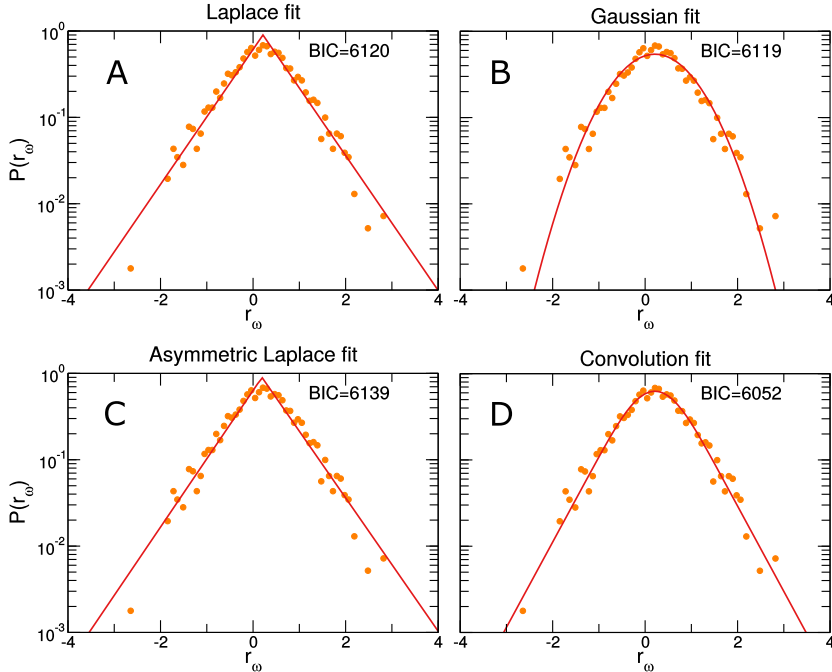


Figure 2.4: **Models for the distributions of weight logarithmic growth rates r_ω for $\Delta t = 1$ and $t = 2007$.** Orange circles correspond to $P(r_\omega)$. Red lines show the maximum likelihood fits. In the top right of each graph we show the BIC for each fit. **(A)** Laplace fit according to Eq. (2.3). **(B)** Gaussian fit according to Eq. (2.4). **(C)** Asymmetrical Laplace fit, according to Eq. (2.5). **(D)** Convolution fit according to Eq. (2.6). We obtain same results for other starting years t and values of Δt .

the intensity of the Gaussian noise also increases with Δt . The only exception is that for $P(r_\omega^0)$, σ_{exp}^ω seems to stop growing for $\Delta t = 3$.

Tent-shaped distributions with exponentially decaying tails are common in the growth of human organizations (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999), and have also been reported in the growth of complex weighted networks (Gautreau et al. 2009). The exponential tails of these distributions imply that fluctuations in connection weights and user strengths are considerably larger than one would expect from a process with Gaussian-like fluctuations.

2.4 Social signatures are stable in the long term

Next, we seek to better understand the evolution of the communication behavior of individual users. Recent results suggest that the way individuals divide their commu-

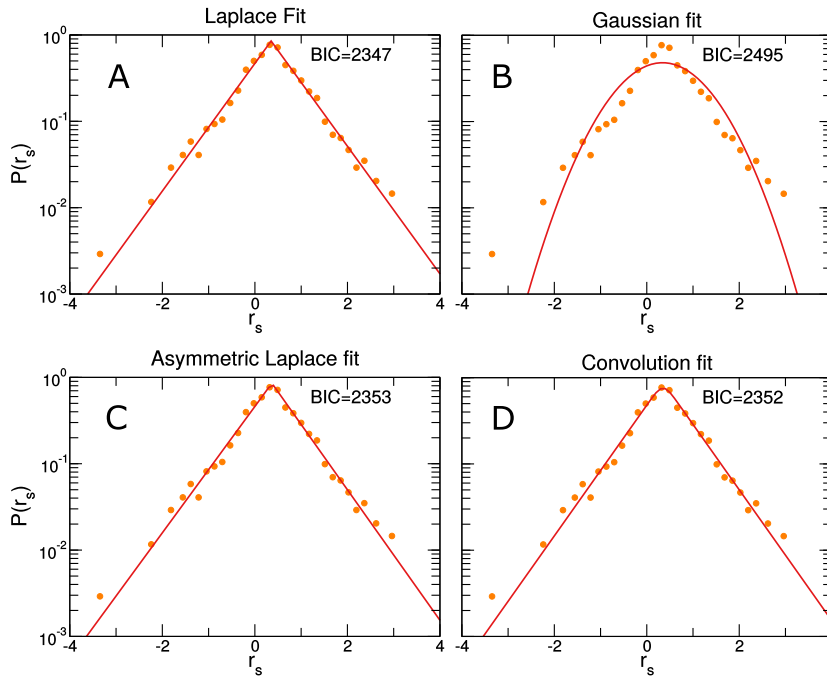


Figure 2.5: **Models for the distributions of strength logarithmic growth rates r_s for $\Delta t = 1$ and $t = 2007$.** Orange circles correspond to $P(r_s)$. Red lines show the maximum likelihood fits. In the top right of each graph we show the BIC for each fit. **(A)** Laplace fit according to Eq. (2.3). **(B)** Gaussian fit according to Eq. (2.4). **(C)** Asymmetrical Laplace fit, according to Eq. (2.5). **(D)** Convolution fit according to Eq. (2.6). We obtain same results for other starting years t and values of Δt .

nication effort among their contacts (their so-called “social signature”) is stable over the period of a few months (Saramäki et al. 2014). This is consistent with the hypothesis that humans have a limited capacity to simultaneously maintain a large number of social interactions (Dunbar 1998). Thus, the users develop different communication strategies of communication as it has been shown that some individuals tend to change their contacts frequently (“explorers”), whereas others tend to maintain contacts (“keepers”) (Miritello et al. 2013).

We investigate whether these differences exist at the scale of years and if individual signatures are stable in the long term. To do so, we investigate three different aspects of the email communication: i) How users divide their communication among their contacts—through a standardized Shannon entropy measure; ii) the turnover of contacts of a given user from year to year; and iii) the fraction of emails sent from a users to pre-existing contacts with respect to the total amount. We found that individuals have social signatures that are stable in the long-term.

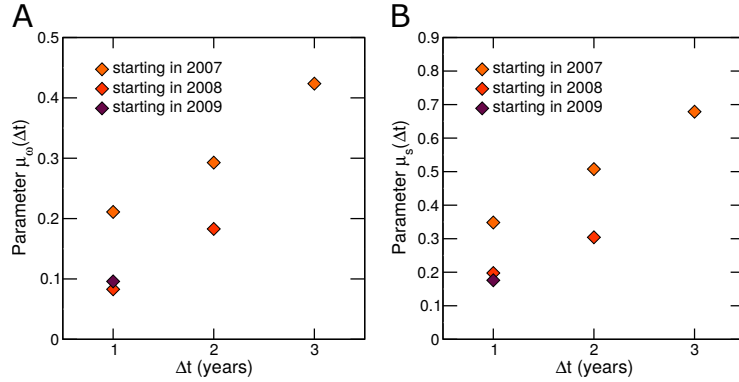


Figure 2.6: **Time evolution of the modes μ for the distribution of logarithmic growth rates.** (A,B) We show $\mu_\omega(t, \Delta t)$ and $\mu_s(t, \Delta t)$ for $t = 2007, 2008, 2009$ and $\Delta t = 1, 2, 3$.

Comparison ($X, \Delta t, (t_1, t_2)$)	Pair	KS Statistic	p-value
$\omega, \Delta t = 1, (2007, 2008)$		0.022	0.45
$\omega, \Delta t = 1, (2008, 2009)$		0.025	0.19
$\omega, \Delta t = 1, (2007, 2009)$		0.037	0.021
$\omega, \Delta t = 2, (2007, 2008)$		0.031	0.18
$s, \Delta t = 1, (2007, 2008)$		0.032	0.63
$s, \Delta t = 1, (2008, 2009)$		0.031	0.61
$s, \Delta t = 1, (2007, 2009)$		0.029	0.71
$s, \Delta t = 2, (2007, 2008)$		0.039	0.42

Table 2.1: Kolmogorov-Smirnov test comparison results. We compare pairs of distributions of centered LGRs for fixed Δt , ($P(r_X^0; t_1, \Delta t), P(r_X^0; t_2, \Delta t)$) for $X = \omega, s$. If the p-value is greater than 0.01, we cannot reject the null hypothesis that both distributions are the same (at a 1% significance level).

2.4.1 Analysis of how individuals distribute their communication

Here, we analyze how individuals distribute their communication activity (their emails) among their contacts. To quantify how evenly distributed emails are among those contacts, we thought about two different measures: the Gini coefficient and the standardized Shannon entropy.

The Gini coefficient 2.7, which is used to measure inequalities in wealth distributions within and across countries, measures the disparity of weights between the different connections of each individual. The Gini coefficient is a well-known measure of dispersion in economy to quantify the inequality of social income in society (Yitzhaki

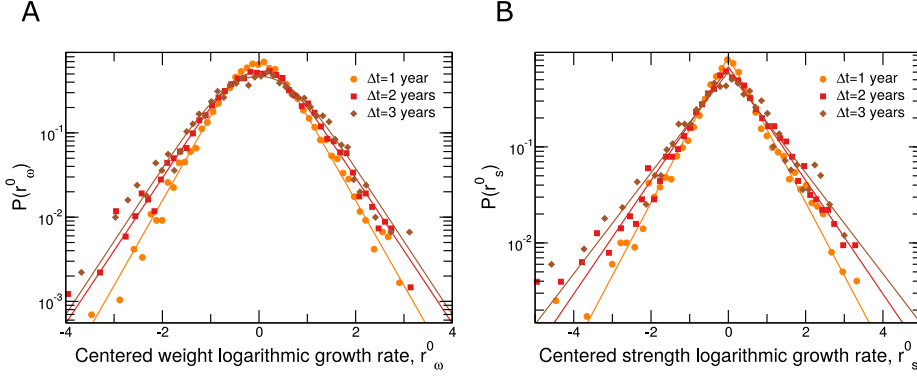


Figure 2.7: **Stability of the distributions of centered logarithmic growth rates.** (A) Distribution of the centered weight logarithmic growth rates $r_\omega^0 = \log(\omega(t+\Delta t)) - \log(\omega(t)) - \mu(t, \Delta t)$ for $\Delta t = 1, 2, 3$ (dots, squares and diamonds, respectively). Lines show fits to the convolution of a Laplace distribution and a Gaussian distributed noise (see Eq. (2.6)) (parameters $\Delta t = 1$: $\sigma_{\text{exp}} = 0.43$, and $\sigma_G = 0.35$, $\Delta t = 2$: $\sigma_{\text{exp}} = 0.50$, and $\sigma_G = 0.47$ and $\Delta t = 3$: $\sigma_{\text{exp}} = 0.50$, and $\sigma_G = 0.60$). Note that as Δt increases the peaks are rounder and the distributions are slightly wider (see Fig. 2.8). For the specific values of the distribution modes $\mu(t, \Delta t)$ see Fig. 2.6. (B) Distribution of centered strength logarithmic growth rates $r_s^0 = \log(s(t+\Delta t)) - \log(s(t)) - \mu(t, \Delta t)$ for $\Delta t = 1, 2, 3$ years (dots, squares and diamonds, respectively). Lines show fits to a Laplace distribution (parameters $\Delta t = 1$: $\sigma_{\text{exp}} = 0.57$, $\Delta t = 2$: $\sigma_{\text{exp}} = 0.74$ and $\Delta t = 3$: $\sigma_{\text{exp}} = 0.83$). Note that as Δt increases the distributions are wider (see Fig. 2.8). For the specific values of the distribution modes $\mu(t, \Delta t)$ see Fig. 2.6.

1979),(Dorfman 1979). In our case, we compute the Gini coefficient as:

$$G_i = \left| 1 - \frac{2}{k_i} \sum_{z=2}^{k_i} X_z^i \right|, \quad (2.7)$$

where k_i is the number of contacts of user i , and \mathbf{X}^i is the cumulative proportion of emails exchanged by user i and her contacts when we order contacts by increasing fraction of exchanged emails. Therefore, $X_z^i = \sum_{c=1}^z f_c$ with f_c the fraction of emails exchanged with contact in position c . If the Gini coefficient is equal to one, then the correspondence is very unequal among the contacts with most of its weight in a single channel. If the Gini coefficient is equal to zero, then, all the channels have the same flow.

The Shannon entropy is another measure to quantify whether the distribution of the flow communication among contacts is even or not. In our case, we define the standardized Shannon entropy so that it takes positive values between zero and one as,

$$S_i = \frac{-\sum_{j=1}^{k_i} \frac{\omega_{ij}}{s_i} \log \frac{\omega_{ij}}{s_i}}{\log k_i}, \quad (2.8)$$

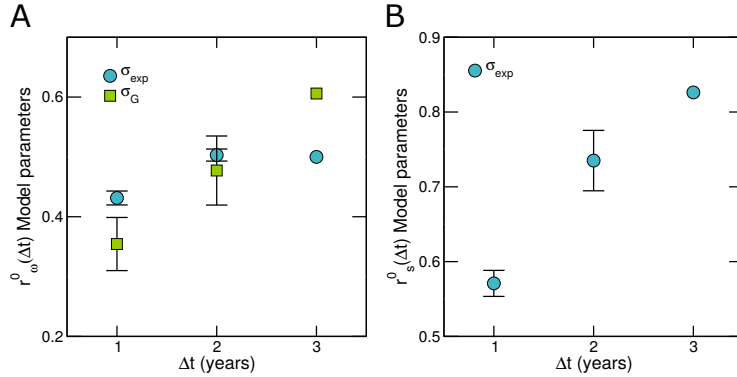


Figure 2.8: **Time evolution of the model parameters for the distribution of logarithmic growth rates.** (A) $P(r_\omega^0)$ model parameters estimated from the maximum likelihood of the convolution of a Laplace distribution and a Gaussian, with two parameters σ_{exp}^ω and σ_G^ω (Eq. 2.6) for $\Delta t = 1, 2, 3$ years. (B) $P(r_s^0)$ model parameters estimated from the maximum likelihood of a Laplace distribution, with parameter σ_{exp}^s (Eq. 2.3) for $\Delta t = 1, 2, 3$ years. The errors are $< 5\%$ in all the cases.

where k_i is the number of contacts of user i . Note that $S_i = 1$ when user i exchanges the same number of emails with all her contacts and $S_i \approx 0$ when she exchanges almost all of her emails with a single contact.

While these two measures are normalized, it is not clear whether they show a systematic trend as we increase the number of contacts. To assess the effect of the number of contacts, for a given number of contacts with values from $c = 3$ to $c = 20$, we measure the average Gini coefficient and the standardized Shannon entropy over $N = 10000$ randomizations of the weights (see Fig.2.9). While the average of the standardized Shannon entropy increases slightly and progressively, with a total change in the average of 0.015 from $c = 3$ to $c = 20$, the Gini coefficient has larger increase for small number of contacts until its arrive to a plateau at 20 contacts, with a change in the average of the Gini coefficient of 0.27 between $c = 3$ and $c = 20$.

Note, that since the majority of users have few contacts (90% of users have less than 20 contacts), the variation of the standardized Shannon entropy in that range is very small. Therefore we use the standardized Shannon entropy for the study of the social signature.

We find that the distribution of standardized entropies is heavily shifted towards high values of S_i (Fig. 2.10A), which implies that most individuals tend to distribute their communication evenly among all their contacts. We also find that the overall distribution of the standardized entropies is stable in time.

To study the stability of each individual's social signature, we measure the difference $\Delta S_i(\Delta t) = S_i(t + \Delta t) - S_i(t)$ for $\Delta t = 1, 2, 3$ years. First we checked that for a fixed value of Δt the distributions of $\Delta S_i(t, \Delta t)$ are stationary (see table 2.4.1), thus in Fig. 2.10B they are pooled together. We find that the distribution of $\Delta S_i(\Delta t)$

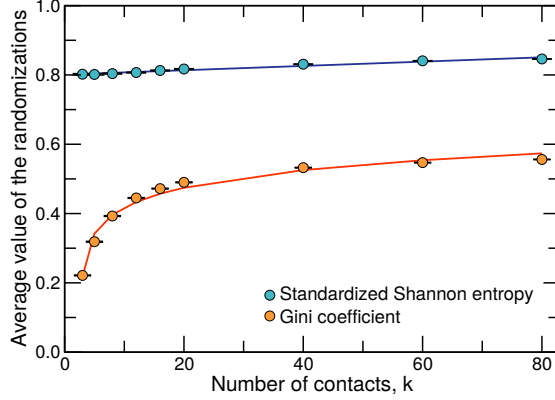


Figure 2.9: **Size effect of the standardized Shannon entropy and the Gini coefficient.** For a fixed number of contacts, we generate $N = 10000$ random samples using link weights from our data. We show the average of the standardized Shannon entropy and the Gini coefficient (blue and orange dots respectively) as a function of the number of contacts. The lines show the best fit of the average: a linear fit for the standardized Shannon entropy (in dark blue, $f(\bar{S}) = 0.80 + 0.0062 \cdot \bar{S}$) and a logarithmic fit for the Gini coefficient (in red, $f(\bar{G}) = 0.28 + 0.66 \cdot \ln(\bar{G} - 2.63)$).

is symmetric and heavily peaked around zero. Therefore since most of the users do not change their social signature during the three year period of our analysis, our results suggest that individual's social signatures are stable in the long term.

Comparison Pair ($X, \Delta t, (t_1, t_2)$)	KS Statistic	p-value
$\Delta S_i, \Delta t = 1, (2007, 2008)$	0.044	0.35
$\Delta S_i, \Delta t = 1, (2008, 2009)$	0.044	0.28
$\Delta S_i, \Delta t = 1, (2007, 2009)$	0.028	0.86
$\Delta S_i, \Delta t = 2, (2007, 2008)$	0.071	0.03

Table 2.2: Kolmogorov-Smirnov test comparison results. We compare pairs of distributions of $\Delta S_i(\Delta t, t)$ for fixed Δt . If the p-value is greater than 0.01, we cannot reject the null hypothesis that both distributions are the same (at a 1% significance level).

To quantify this more precisely, we compare the absolute change of a user's standardized entropy $|\Delta S_i(\Delta t)|_{\text{self}} = |S_i(t + \Delta t) - S_i(t)|$ to the typical absolute difference of entropies between individuals $|\Delta S_{ij}|_{\text{ref}} = |S_i(t) - S_j(t)|, \forall j \neq i$ (Fig. 2.10C). We observe that the variation of the social signature of a user in time is typically much smaller (even when $\Delta t = 3$ years) than the variation between individuals, confirming that the social signature is a trait of users that persists even during periods of several years. In fact, by extrapolating the values of $|\Delta S_i(\Delta t)|_{\text{self}}$, we estimate that individual social signatures may be persistent for roughly eight years.

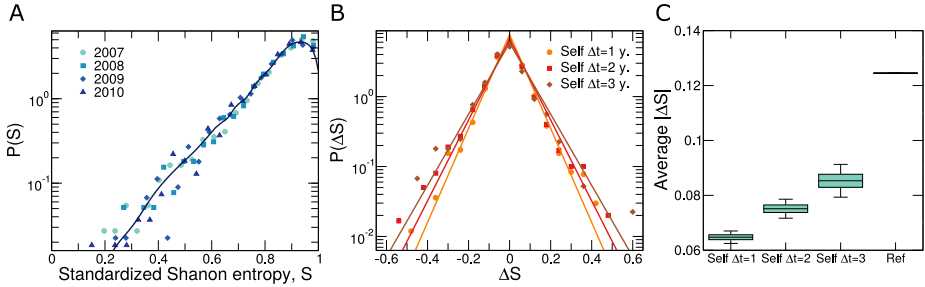


Figure 2.10: **Stability of social signatures.** (A) Distribution of the standardized Shannon entropy S_i (see text) for users in the period 2007–2010. Entropy quantifies the extent to which and individual’s communication efforts are distributed among her contacts, so that $S_i = 1$ when user i exchanges the same number of emails with all her contacts and $S_i \approx 0$ when she exchanges almost all of her emails with a single contact. Distributions for all years collapse onto a single curve. The line shows a kernel density estimation of the four yearly datasets pooled together. (B) Distributions of the change of individual standardized Shannon entropy $\Delta S_i(\Delta t) = S_i(t + \Delta t) - S_i(t)$, $\forall i$ for $\Delta t = 1, 2, 3$ years (dots, squares and diamonds, respectively). The lines show the Laplace best fits based on BIC for the three distributions ($\Delta t = 1 \sigma = 0.065$; $\Delta t = 2 \sigma = 0.075$; and $\Delta t = 3 \sigma = 0.085$). (C) Comparison between the absolute difference in individual social signatures $|\Delta S_i(\Delta t)|_{\text{self}} = |S_i(t + \Delta t) - S_i(t)|$ and the typical absolute difference of entropies between individuals $|\Delta S_{ij}|_{\text{ref}} = |S_i(t) - S_j(t)|$. The boxplot shows unambiguously that users have stable social signatures.

2.4.2 Analysis of the contacts turnover of users

A related question to the stability of the social signature is that of whether users tend to keep the same contacts over time or not. Specifically, we measure the fraction f_{k_i} of contacts with whom user i exchanged emails in years t and $t - 1$ compared to the total number of contacts k_i during year $t - 1$

$$f_{k_i}(t) = \frac{k_i(t) \cap k_i(t-1)}{k_i(t)}. \quad (2.9)$$

Therefore, if $f_{k_i}(t) = 1$, user i has maintained all her contacts from the previous year (regardless of the number of emails that has exchanged with each one of them), whereas $f_{k_i}(t) = 0$, user i has changed all her contacts.

The distribution of f_k (Fig. 2.11A) indicates that most individuals tend to maintain the majority of their contacts from year to year. The several peaks of the distribution are due to the discrete number of values that fractions can take on those users with few contacts (52% of users have less than 5 contacts). Although the distribution is irregular, the mass of the distributions is slightly shifted to higher values of f_k , with a 62% of the mass by $f_k > 0.5$.

To study the stability of each individual’s turnover in the long term, we measure the change $\Delta f_{k_i}(\Delta t) = f_{k_i}(t + \Delta t) - f_{k_i}(t)$ at $\Delta t = 1$ year and $\Delta t = 2$ years (Fig. 2.11B). The distributions are stationary for any fixed value of Δt ($KS(\Delta f_k(t) =$

08, $\Delta t = 1$), $\Delta f_k(t = 09, \Delta t = 1)$) = 0.041, p-value = 0.36; as the p-value is greater than 0.01, we cannot reject the null hypothesis that both distributions are the same at a 1% significance level). We also observe that most users do not change their turnover from year to year (the mode is at $\Delta f_k(\Delta t) = 0$). However, 11% of the individuals change their communication strategy by $|\Delta f_k(\Delta t)| > 0.5$.

Despite this variability, we find that, on average, an individual's turnover is stable in the long term (Fig. 2.11C). In particular, we compare the absolute individual change $|\Delta f_{k_i}(\Delta t)|_{\text{self}} = |f_{k_i}(t + \Delta t) - f_{k_i}(t)|$ with the typical absolute difference between individuals $|\Delta f_{k_{ij}}|_{\text{ref}} = |f_{k_i}(t) - f_{k_j}(t)|$, $\forall j \neq i$. We observe that the yearly variation of an individual's communication strategy is typically much smaller (even when $\Delta t = 2$ years) than the variation between individuals, confirming the existence of persistent turnover signatures even at the scale of several years.

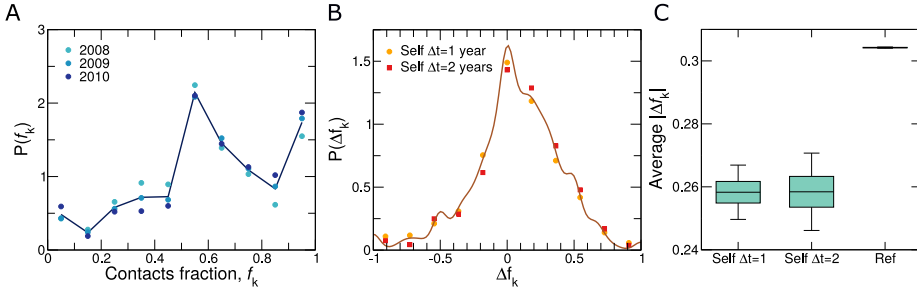


Figure 2.11: **Variability of individual turnover of contacts.** (A) Distributions of the fraction of contacts that are preserved in two consecutive years $f_{k_i}(t)$ (see Eq.2.9). The line shows the distribution for the aggregate of the three datasets. Note that the distributions are stable in time. (B) Distributions of $\Delta f_{k_i}(\Delta t) = f_{k_i}(t + \Delta t) - f_{k_i}(t)$, the change in the fraction of preserved users, for $\Delta t = 1, 2$ years (dots and squares respectively). Note that the distributions are stable in time. The line shows the smooth kernel distribution for the two distributions pooled together. (C) Boxplot of the average over all the users of the absolute differences in the fraction of preserved contacts of each user (self) $|\Delta f_{k_i}(\Delta t)|_{\text{self}} = |f_{k_i}(t + \Delta t) - f_{k_i}(t)|$ for $\Delta t = 1, 2$ years compared to the difference in the fraction of preserved contacts between a random pair of users (reference) $|\Delta f_{k_{ij}}|_{\text{ref}} = |f_{k_i}(t) - f_{k_j}(t)|$. Note that users have stable turnover since the difference in the fraction of preserved contacts for an individual is small compared to the average difference between pairs of users.

2.4.3 Analysis of the fraction of emails to pre-existing contacts

Once we know how the users change their contacts from year to year, we analyzed which is the actual impact of the new acquaintances with respect with the old ones. For this aim, we study how the emails are distributed among the old and new contacts. We define the fraction $f_{e_i}(t)$ of all emails exchanged by user i in year t (out of the total $s_i(t)$) with pre-existing contacts—that is users with whom user i had also exchanged emails during the previous year, $t - 1$. Therefore, $f_{e_i}(t) = 1$ means user i exchanged

all her emails in year t with pre-existing contacts, whereas $f_{e_i}(t) = 0$ means that user i only exchanged emails with new contacts.

The distribution of f_{e_i} (Fig. 2.12A) shows that most individuals are social keepers. Indeed, the mode of the distribution is around $f_{e_i}(t) = 0.9$, and 58% of the users exchange more than 75% of their emails with pre-existing contacts. Still, a non-negligible 17% of the individuals exchange more than half of their emails in one year with new contacts. Our findings thus confirm that, even at the scale of years, there is a variety of communication strategies (Miritello et al. 2013).

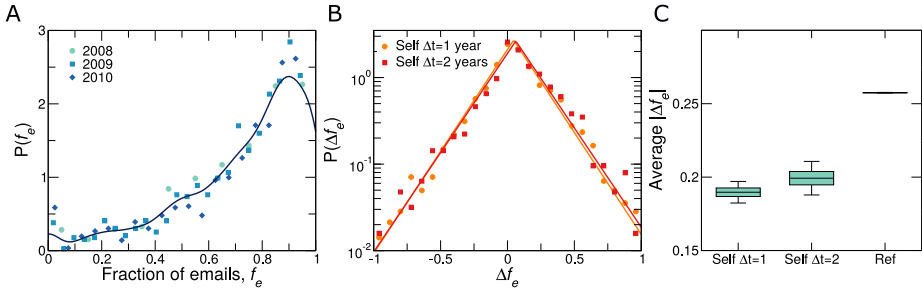


Figure 2.12: Stability of individual fraction of emails to pre-existing contacts. (A) Distribution of the fraction of emails sent by users to pre-existing contacts f_{e_i} (see text). The line shows the kernel density estimation of the three yearly datasets pooled together. Most users exchange most of their emails with pre-existing contacts, with the maximum at $f_{e_i}^{max} = 0.90$. (B) Distribution of the change of f_{e_i} , $\Delta f_{e_i}(\Delta t) = f_{e_i}(t + \Delta t) - f_{e_i}(t)$ for $\Delta t = 1, 2$ years (dots and squares, respectively). The lines show the Laplace best fits based on BIC for the two distributions ($p(\Delta f_{e_i}) \sim \exp(-|\Delta f_{e_i} - \mu|/\sigma)$; $\Delta t = 1$ $\sigma = 0.18$ $\mu = 0.046$; and $\Delta t = 2$ $\sigma = 0.19$ $\mu = 0.062$). Most of the users keep the number of emails sent to pre-existing contacts constant in time, and the distributions are quite stable in time despite a slight shift towards larger changes for larger Δt . (C) Comparison between yearly absolute individual change in the fraction of emails sent to pre-existing contacts $|\Delta f_{e_i}(\Delta t)|_{self}$ and the typical differences between users $|\Delta f_{e_{ij}}|_{ref} = |f_{e_i}(t) - f_{e_j}(t)|$, $\forall j \neq i$. The boxplot shows unambiguously that individual users have a stable signature over time.

To study the stability of each individual's strategy in the long term, we measure the change $\Delta f_{e_i}(\Delta t) = f_{e_i}(t + \Delta t) - f_{e_i}(t)$ at $\Delta t = 1$ year and $\Delta t = 2$ years (Fig. 2.12B). First, we find that the distributions are stationary for any fixed value of Δt ($KS(\Delta f_{e_i}(t = 08, \Delta t = 1), \Delta f_{e_i}(t = 09, \Delta t = 1)) = 0.071$, p-value = 0.023; as the p-value is greater than 0.01, we cannot reject the null hypothesis that both distributions are the same at a 1% significance level). From the distributions, we also observe that most users do not change substantially their communication strategy from year to year. However, 7% of the individuals change their fraction of emails to pre-existing contacts by $|\Delta f_{e_i}(\Delta t)| > 0.5$, and a small fraction of individuals even change from one end to the other of the f_e spectrum.

Despite this variability, we find that, on average, an individual's communication strategy is stable in the long run (Fig. 2.12C). In particular, we compare the absolute

individual change $|\Delta f_{e_i}(\Delta t)|_{\text{self}} = |f_{e_i}(t + \Delta t) - f_{e_i}(t)|$ with the typical absolute difference between individuals $|\Delta f_{e_{ij}}(t)|_{\text{ref}} = |f_{e_i}(t) - f_{e_j}(t)|$, $\forall j \neq i$ (Saramäki et al. 2014). We observe that the yearly variation of a user's f_e is typically much smaller (even when $\Delta t = 2$ years) than the variation between individuals, confirming the existence of individual signatures on the fraction of emails to pre-existing contacts persistent even at the scale of several years. By extrapolating the values of $|\Delta f_{e_i}(\Delta t)|_{\text{self}}$ as before, we estimate that individual signatures may persist for around seven years.

2.5 Discussion

We have shown that the long-term macro-evolution of email networks follows well-defined distributions, characterized by exponentially decaying log-variations of the weight of social ties and of individuals' social strength. Therefore, the intricate processes of tie formation and decay at the micro-level give rise to macroscopic evolution patterns that are similar to those observed in other complex networks (such as air-transportation or financial networks (Gautreau et al. 2009)), as well as in the growth and decay of human organizations (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999).

Remarkably, together with these statistical regularities, we also observe that individuals have long-lasting social signatures in how they distribute their communication, in the turnover of contacts and in the fraction of emails to pre-existing contacts, confirming the social signature found by (Saramäki et al. 2014) and communication strategies (Miritello et al. 2013; Miritello et al. 2013) in phone call networks. Reconciling the universality of the macroscopic evolutionary patterns with the importance of the psychological/microscopic processes should be one of the central aims of future studies about the evolution of social networks.

Predicting collective and individual social behavior

3.1 Introduction

There are many social behaviors of individuals we cannot predict due to their erratic and subjective nature. However, a plausible hypothesis is that if individuals of social systems develop stable patterns of behavior over time, then that behavior could be predictable. For instance, human trajectories show a high degree of temporal and spatial regularity, where individuals present significant probability to return to high frequency locations (González et al. 2008); also judges' votes in a Court can be predicted based on previous voting patterns (Guimerà and Sales-Pardo 2011).

In Chapter 2 we have shown collective and individual patterns on email correspondence. In this section we explore the possibility of predicting such a behaviors based on these patterns. First we analyze the predictability of the logarithmic growth rates (LGRs) for both weights and strengths based on the correlations with several network features and using well-performing machine learning algorithms. To find significant correlations has played a larger role in statistical inference for data dimension reduction and for prediction. However, finding variables significantly correlated with the variable to predict does not necessarily lead to predictive power (Lo et al. 2015). For the logarithmic growth rate, we find that despite strong correlation between network features, the LGRs are highly unpredictable. At the same time, we analyze to what extent the social signature could make users indistinguishable. We find that although it

is impossible to uniquely reidentify users, we can rank the possible candidates with a maximum recall of 78%.

3.2 Logarithmic growth rates are largely unpredictable despite significant correlations

The fact that long-term growth rates follow well-defined distributions raises the question of whether it is possible to quantitatively predict the evolution of the network. Here, we study whether there are long-term trends in the logarithmic growth rates. First we analyze the correlations with a number of network features we suspect that can be relevant to the evolution of the LGRs. Then, to quantify their predicting power, we perform leave-one-out experiments with the most relevant features and using a Random Forest regressor (Breiman 2001). The Random forest (RF) algorithm is a popular machine learning method consisting of a multitude of decision trees; the strengths of the RF approach are that it does not overfit, it is robust to noise and it provides indices of variable importance.

For this analysis we use the uncentered logarithmic growth rates. In the previous chapter (see Fig. 2.6) we have shown that the LGRs distribution are not stationary, however we can not use this information in our predictive analysis since we do not know the future distributions of growths. Additional, we know that prediction errors are in any case larger than the shift of the mode of the distributions. Therefore we use r_ω and r_s in the prediction analysis.

3.2.1 Predicting weights' LGRs $r_\omega(t + 1)$

To assess the predictability of $r_\omega(t + 1)$ we analyzed the correlation with a number of network features that we could measure at time t . We choose an array of network features that we thought could bear a relationship with the evolution of communication weights. Specifically, for each edge (i, j) we measured:

- $\omega_{ij}(t)$: the undirected total weight of the edge;
- the betweenness centrality of the edge: that is the sum of the fraction of all-pairs shortest paths that pass through the edge.
- the relative weight of the edge: $\bar{r}_{ij} = (\omega_{ij} \cdot k_i \cdot k_j) / (s_i \cdot s_j)$, where k_i is the degree of i and s_i is the strength of i ;
- $r_\omega^{ij}(t)$: the weight logarithmic growth rate of the edge;
- the Jaccard index of the edge $J_{ij} = \frac{|\text{neigh}(j) \cap \text{neigh}(i)|}{|\text{neigh}(j) \cup \text{neigh}(i)|}$;
- the maximum node strength $\max\{s_i, s_j\}$;
- the maximum node degree $\max\{k_i, k_j\}$;

- the maximum node betweenness: the higher node betweenness of the two nodes which form the edge. The node betweenness is the sum of the fraction of all-pairs shortest paths that pass through the node $bet(i)$;
- the maximum node clustering: the higher clustering coefficient of the two nodes which form the edge. The clustering coefficient of a node is the fraction of possible triangles through that node that exist,

$$c_i = \frac{2T(i)}{k_i(k_i - 1)}, \quad (3.1)$$

where $T(i)$ is the number of triangles through node i and k_i is the degree of i ;

- the absolute difference in node strength $|s_i - s_j|$;
- the absolute difference in node betweenness $|bet(i) - bet(j)|$;
- the absolute difference in node clustering $|c_i - c_j|$;

Figure 3.1 shows the density plots of $r_s(t + 1)$ versus the different features above ordered from higher to lower significance. First of all, we observe that correlations are significant for about half of the features we analyze. However, it is obvious that even for the most significantly correlated feature $\omega(t)$, the variability of $r_\omega(t + 1)$ for a fixed value of $\omega(t)$ is too large for this feature to produce accurate predictions (see below in Fig. 3.3A) as indicated by the rather modest value of Spearman's $\rho = -0.24$. Also the correlation with $\omega(t)$ is negative, which indicates that small values of connection weight grow faster than large values, also because negative values of weights are not allowed. To analyze the existence of a long-term trends in the evolution of the LGRs, we measure the correlation between the logarithmic growth rate in one year and the logarithmic growth rate the following year. We find that the correlation is significant but negative for weight logarithmic grow rates (Spearman's $\rho = -0.16$, $p = 1.5 \cdot 10^{-27}$).

3.2.2 Predicting the strengts' LGRs $r_s(t + 1)$

To assess the predictability of $r_s(t + 1)$ we analyzed the correlation with a number of network features that we could measure at time t . We choose an array of network features that we though could bear a relationship with the evolution of node communication strengths. Specifically, for each node i we measured:

- $s_i(t)$: total strength of the node;
- the eigen vector centrality: It is the centrality for a node based on the centrality of its neighbours. The eigenvector centrality for node i is

$$\mathbf{Ax} = \lambda\mathbf{x}$$

where A is the adjacency matrix of the graph G with eigenvalue λ . By virtue of the Perron–Frobenius theorem, there is a unique and positive solution if λ is the largest eigenvalue associated with the eigenvector of the adjacency matrix A ;

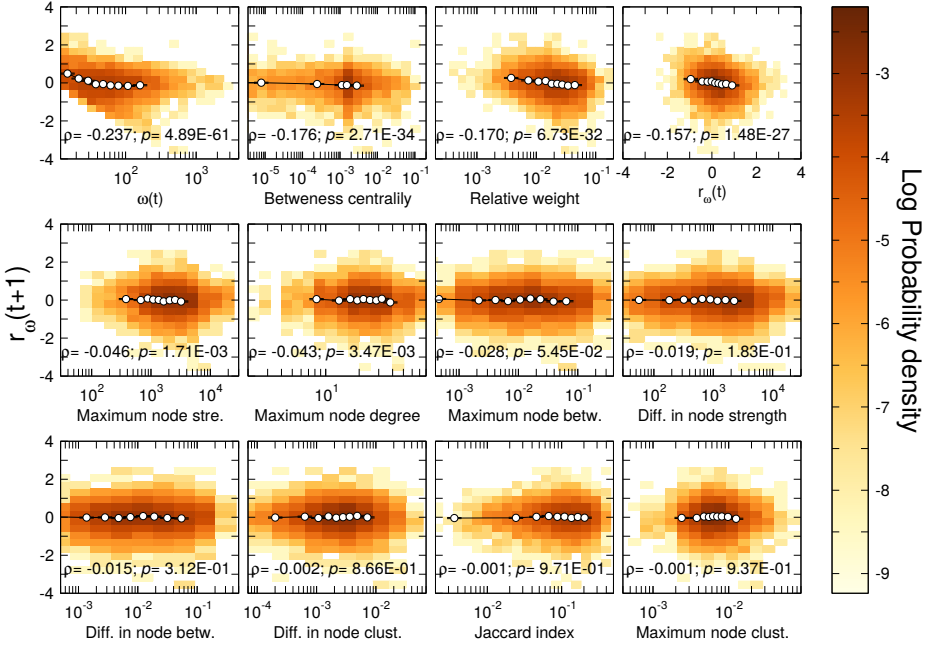


Figure 3.1: **Long-term trends on the weight logarithmic growth rates.** Density plot of the weight logarithmic growth rate $r_{\omega}(t + \Delta t = 1)$ as a function of the 12 network features mentioned in the text. The lines correspond to the mean and the error of the mean in each bin along the x axis. We show the Spearman's ρ and the significance of the correlation at the bottom of each graph.

- k_i : degree of the node;
- $r_s(t)$: the strength logarithmic growth rate of the node;
- the clustering of node i : The clustering coefficient of a node is the fraction of possible triangles through that node that exist (see Eq. 3.1);
- the size of the largest clique containing node i : a clique in networks is a subset of nodes where all nodes are connected among them, then from a given node that could be in several cliques, we take the larger one;
- the betweenness centrality of node i : it is the sum of the fraction of all-pairs shortest paths that pass through the node;
- the closeness centrality of node i : it is the reciprocal of the sum of the shortest path distances from i to all $N-1$ other nodes. Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of

minimum possible distances $N-1$.

$$C(u) = \frac{N-1}{\sum_{j=1}^{N-1} d(j, i)}, \quad (3.2)$$

where $d(j, i)$ is the shortest-path distance between j and i , and N is the number of nodes in the graph.;

- the load centrality of node i : it is the fraction of all shortest paths that connect any pair of nodes in the network that pass through i .

Figure 3.2 shows the density plots of $r_s(t+1)$ versus the different features above ordered from higher to lower significance. As for the weights, we observe that correlations are significant for about half of the features we analyze. In the same way, even for the most significantly correlated feature $s(t)$, the variability of $r_s(t+1)$ for a fixed value of $s(t)$ is too large for this feature to produce accurate predictions (see below in Fig. 3.3B). Again, the correlation with $s(t)$ is negative, which indicates that small values of user strength grow faster than large values, also because negative values of strengths are not allowed. Remarkably, we find that there is not a significant correlation between $r_s(t+1)$ and $r_s(t)$.

3.2.3 Leave-one-out experiments

We find that the network properties at time t that are most correlated with the logarithmic growth rates $r_\omega(t+1)$ and $r_s(t+1)$ are the connection weight and the user strength, respectively (see Figs. 3.1 and 3.2). In any case, despite the significance of these correlations, the high variability of $r_\omega(t+1)$ and $r_s(t+1)$ for fixed values of $\omega(t)$ and $s(t)$, respectively, raises the question of whether the correlations can be used reliably to predict the evolution of the network.

To quantify the predictive power of these variables, we carry out leave-one-out experiments to predict logarithmic growth rates $r_\omega(t+1)$ and $r_s(t+1)$ from network properties at time t (Fig. 3.3). Consider a dataset (x, r) in which x is the network feature and r is the corresponding logarithmic growth rate. In general, we find that we can mathematically model the dependence of r in x . To assess the predictability of logarithmic growth rates from network features at time t , we perform leave-one-out experiments for selected network features. For each point in our dataset (x_i, r_i) , we construct a new training dataset in which we remove this point. Then we train our model (that is, we estimate the model parameters) with the training dataset. Finally, we obtain a prediction p_i of r_i from the trained model using x_i as our input. To estimate the accuracy of the predictions, we compute the mean squared error (MSE), that is $\text{MSE} = \frac{1}{N} \sum_{i=1}^N (p_i - r_i)^2$. The sample sizes are $N_\omega = 4,721$ and $N_s = 2,013$ for the prediction of $r_\omega(t+1)$ and $r_s(t+1)$, respectively.

The features we consider are:

- the most significantly correlated features $x = \omega(t), s(t)$, for which we assume that $r = A \exp(-x \cdot B) + C$;

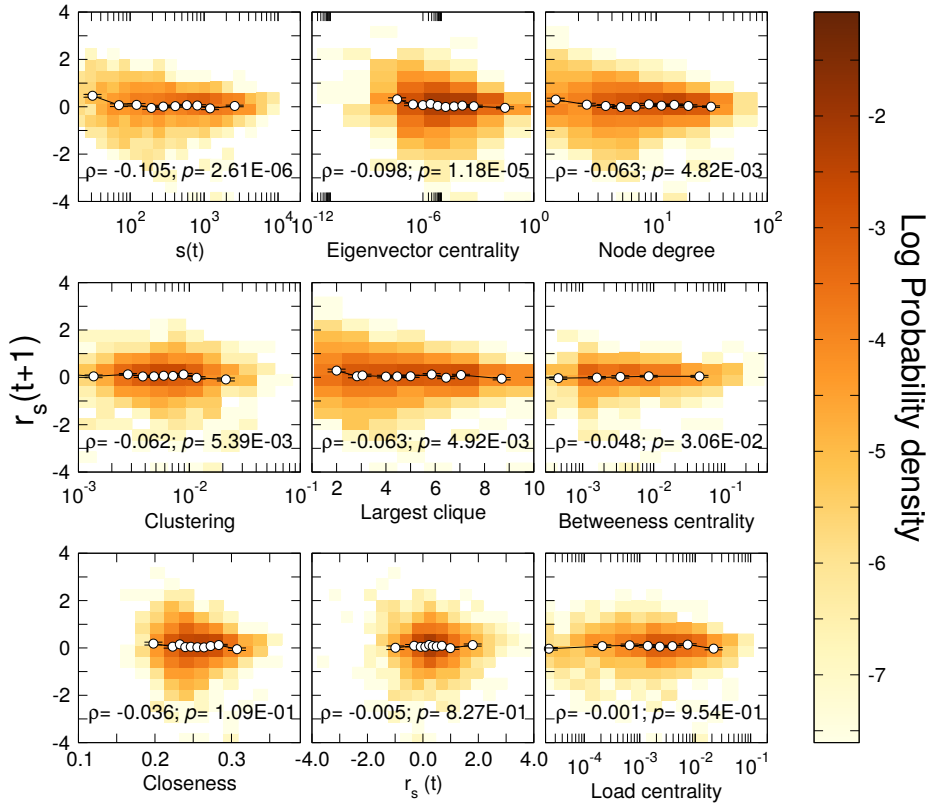


Figure 3.2: **Long-term trends on the strength logarithmic growth rates.** Density plot of the strength logarithmic growth rate $r_s(t + \Delta t = 1)$ as a function of the 9 network features mentioned in the text. The lines correspond to the mean and the error of the mean in each bin along the x axis. We show the Spearman's ρ and the significance of the correlation at the bottom of each graph.

- the previous value of the variable we want to predict $x = r_\omega(t + 1), r_s(t + 1)$ for which we assume $r = A \cdot x + B$;
- the mode of the logarithmic growth rate distributions μ_ω, μ_s which are constants (note that as we show in Fig. 2.6, mean growths are very close to zero).

Finally, we perform leave-one-out experiments using the Random Forest (Breiman 2001). For each training dataset, Random Forest uses *all* the features for edges/nodes we have listed previously as inputs to train the algorithm and produce a prediction for the data point not present in the training data set. Our results show that using the Random Forest does not yield significantly better predictions than using using the most correlated features, $\omega(t)$ and $s(t)$ for $r_\omega(t + 1)$ and $r_s(t + 1)$, respectively.

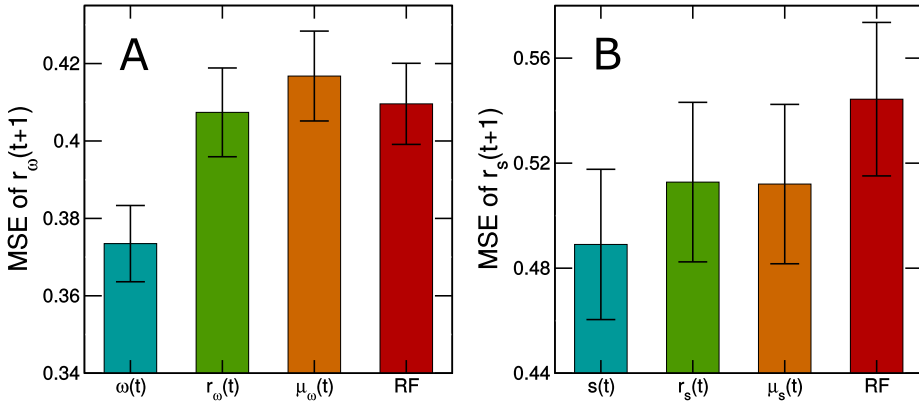


Figure 3.3: **Predictability of logarithmic growth rates for connection weight $r_{\omega}(t + 1)$ and user strength $r_s(t + 1)$.** (A, B) Root mean squared error (MSE) of the predictions of the logarithmic growth rates at time $t + 1$ obtained from leave-one-out experiments. As predictors, we use: (A) $\omega(t)$, $r_{\omega}(t)$, and $\mu_{\omega}(t)$ (see Eq. 2.6); (B) $s(t)$, $r_s(t)$, and $\mu_s(t)$ (see Eq. 2.3). Additionally, in both cases we try to predict the logarithmic growth rate using a Random Forest regressor. Note that a simple approach (i.e. considering the weight/strength at time t) performs significantly better than a well-performing machine learning algorithm such as the Random Forest. In any case, and despite being the most predictive, weight/strength at time t only provide moderate improvements over predictions made using the mean value μ_{ω} for all connections and μ_s for all users.

We find that using the Random Forest does not yield significantly better predictions than using the average expected growth for all predictions. Using the most correlated variables $\omega(t)$ and $s(t)$ for $r_{\omega}(t + 1)$ and $r_s(t + 1)$ respectively, only shows a modest improvement (Fig. 3.3). Our results therefore suggest that the existence of correlations is not enough to build a satisfactory predictive model for the logarithmic growth rates (and that black box methods like Random Forests may, in fact, be even less appropriate).

3.3 Reidentifying users based on their social signature

In this section we explore the possibility to reidentify users based on their past activity. From the previous chapter, we know that each user in the email network has her characteristic communication activity, which defines her social signature. So far we know that this signature is stable in time, here we analyze how different they are among them through the challenge of personal reidentification. Recent studies point out that it is possible to reidentify users in anonymized datasets by adding relatively few outside information. For instance, in a mobile phone dataset where it is also known the location, four spatio-temporal points are enough to uniquely identify 95% of the individuals (de Montjoye et al. 2013); also using credit card records it has been shown

that four spatio-temporal points are enough to uniquely reidentify 90% of individuals (de Montjoye et al. 2015). Our email dataset is limited strictly to the email correspondence of anonymized users, so we do not have an outside source of information to uniquely reidentify the users. In this context, we study how reidentifiable users are based only on their correspondence.

3.3.1 Scores definition

To the reidentification of the users we use two communication features defined in the last chapter: the standardized Shannon entropy $S_i(t)$ (see Eq. 2.8) and the individual strength $s_i(t) = \sum_j \omega_{ij}(t)$.

We hide the last year of communication (2010), and we reidentify them considering the communication features from the pool of candidates from 2007 to 2009. There are other variables, as the turnover of contacts f_{k_i} or the fraction of emails to old contacts f_{e_i} , but to define them we need to hide two years of data (see below in subsection 3.3.2).

Specifically, for each pair (i, j) of hidden and candidate users respectively, we define a score base on the probability of change of standardized Shannon entropy $P(\Delta S_i(\Delta t))$ (see Fig. 2.10B) and the probability distribution of the strength's centered LGR $P(r_{s_i}^0(\Delta t))$ (see Fig. 2.7B) for $\Delta t = 1, 2, 3$ years. These individual measures of change are well-defined by analytical expression which best parameters are fitted in the last chapter.

For the reidentificaton, we consider two different cases that could be defined depending on the pool of candidates: case A) using as candidates active during the four years of study; and case B) using as candidates all users that at least were active during one year between 2007 and 2009. We build the simpler model possible, not taking into account the correlations between the variables. Therefore for case A, where candidates are users active the four years, the scores for a pairs (i, j) of hidden and candidate users are defined as:

- score based on the standardized Shannon entropy

$$I_{ij}^{S,A}(t = 2010) = P(\Delta S_{ij}(\Delta t = 1)) \cdot P(\Delta S_{ij}(\Delta t = 2)) \cdot P(\Delta S_{ij}(\Delta t = 3)); \quad (3.3)$$

- score based on the individual strength

$$I_{ij}^{r,A}(t = 2010) = P(r_{ij}^0(\Delta t = 1)) \cdot P(r_{ij}^0(\Delta t = 2)) \cdot P(r_{ij}^0(\Delta t = 3)); \quad (3.4)$$

- score using both the standardized Shannon entropy and the strength

$$I_{ij}^{both,A}(t = 2010) = I_{ij}^{S,A}(t = 2010) \cdot I_{ij}^{r,A}(t = 2010). \quad (3.5)$$

Where $\Delta S_{ij}(\Delta t)$ and r_{ij}^0 are the changes in the variables between hidden user i and the candidate j defined as: $\Delta S_{ij}(\Delta t) = S_i(t) - S_j(t - \Delta t)$ and $r_{ij}^0 = \log(s_i(t)) -$

$\log(s_j(t - \Delta t) - \mu(t) + \mu(t - \Delta t))$; where $\mu(t)$ would be the average growth in year t as $\mu(t) = \frac{1}{N} \sum_i^N \log(s_i(t))$.

In case B, where the candidates are all users active at least one year of the study, the scores are (notation $i \sim t \equiv i$ active in t and $i \not\sim t \equiv i$ inactive in t):

- score based on the standardized Shannon entropy

$$I_{ij}^{S,B}(t = 2010) = \prod_{t; j \sim t} P(\Delta S_{ij}(\Delta t = 2010 - t)) \cdot \prod_{t; j \not\sim t} (1 - P_{act}(t)); \quad (3.6)$$

- score based on the individual strength

$$I_{ij}^{r,B}(t = 2010) = \prod_{t; j \sim t} P(r_{ij}^0(\Delta t = 2010 - t)) \cdot \prod_{t; j \not\sim t} (1 - P_{act}(t)); \quad (3.7)$$

- score using both the standardized Shannon entropy and the strength

$$I_{ij}^{both,B}(t = 2010) = I_{ij}^{S,B}(t = 2010) \cdot I_{ij}^{r,B}(t = 2010). \quad (3.8)$$

Where the probability of an active user in 2010 to remain active in t is,

$$P_{act}(t) = \frac{\#active\ users(2010) \cap \#active\ users(t)}{\#active\ users(2010)}.$$

3.3.2 Ranking and recall

We use the scores to rank candidates in both cases. If the selected variables are meaningful, we expect the true candidate to have a high score. To assess ranking accuracy we compute the recall as the fraction of candidates that have a lower score than the true candidate. For a perfect classification the recall would be 1, and for meaningless variables the expected recall would be 0.5.

Figure 3.4 shows that, as expected, the recall is better for the case A, where the candidates are active the your years, than for case B, where to be a candidate is enough to be active during one year. This is because of two main reason: i) in case A there is more information and the same amount of information for all the candidates, then when comparing the scores it is clearly a fair comparison; in case B some candidates (the once that are active more years) have more information than other candidates –even thought we compensate it with the probability of being inactive; ii) the pool of possible candidates is higher for the case B than for the case A, then it is an easier problem by definiton. Also we find that the addition of information from the two variables gives better recall than the best one alone, with an average recall of 0.78 for case A and 0.67 for case B. The increase of the recall by using both variable is remarkably for case B, where the strength variable is nearly meaningless (with recall of 0.51).

We also perform the study for the email fraction f_{e_i} , hiding the two last years (2009 and 2010) of communication and reidentificating those users from the other two years

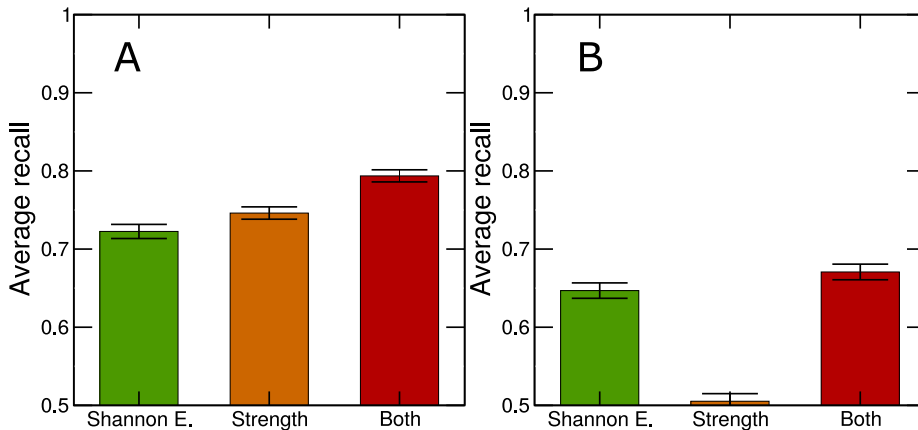


Figure 3.4: **Users reidentification recall using users' Shannon entropy and strength as rank ordering variables.** (A) We show results for the scores considering the standardized Shannon entropy, the strength and both. The average recall corresponds to those candidates that were active during the four years of the study (case A). The recall should be 1 for perfect ranking and 0.5 if the variable is meaningless. Bar colors represent the recall in reidentifying users using standardized Shannon entropy S_i in green (Eqs. 3.3), the individual strength s_i in orange (Eqs. 3.4) and using both variables as they were independent in red (Eqs. 3.5). (B) We show results for the scores considering the standardized Shannon entropy, the strength and both. The average recall corresponds to those candidates that were active at least during one year (case B). The recall should be 1 for perfect ranking and 0.5 if the variable is meaningless. Bar colors represent the recall in reidentifying users using standardized Shannon entropy S_i in green (Eqs. 3.6), the individual strength s_i in orange (Eqs. 3.7) and using both variables as they were independent in red (Eqs. 3.8).

(2007 and 2008). In this case all candidates should be active the four years of study. The average recall in this case is 0.53 ± 0.011 . We have not performed the same analysis for the turnover f_{k_i} , because the distribution of the change $P(\Delta f_{k_i})$ is irregular due to the discrete nature of the variable (see Fig 2.11).

3.4 Discussion

Our results suggest that the email correspondence is highly unpredictable beside the strong correlations between the logarithmic growth rates r_ω and r_s and the different networks features. We observed that the correlations between the LGRs one year and the next one are not significant for strength logarithmic growth rates, and significant but negative for weight logarithmic growth rates. Therefore, we do not observe any long-term trends in the evolution the network. Regarding the predictability, we found that using a black box method such as Random Forest does not perform better than using the average expected growth for all predictions. The best approach, even with very

modest results, is achieved by using the most correlated variables $\omega(t)$ and $s(t)$ for $r_\omega(t+1)$ and $r_s(t+1)$ respectively.

On the other hand, we found that the individual standardized Shannon entropy and strengths could be used to distinguish users among them— with an maximum average recall of 0.78 for the combination of both. The distinguishability is possible thanks to the exponential decay of the changes of both variables, which implies that most of the users do not change their communication strategies in time beside the high variability (Figs. 2.10B and 2.7B for Shannon entropy and strengths respectively). Also of importance is that the combination of different sources of information, in this case the Shannon entropy and the strength, improves significantly the performance in terms of average recall; even when one of them is nearly uninformative.

4

Social recommendation using mixed-membership stochastic block models

4.1 Introduction

Recommender systems have become very common with the ever increasing amounts of different options available through online platforms, modeling and predicting individual preferences (e.g. on movies, books, news, search queries, social tags, items in general). Good predictions enable us to improve the advice to users and reflect a better understanding of the socio-psychological processes that determine those preferences. Several strategies have already been implemented for making recommendations, the most straightforward strategies are based on demographics of the users, overall top selling items, past buying habit of users, etc. to guess preferences of users on items. But collaborative filtering (CF) is the most successful recommendation technique to date. The basic idea of CF algorithms is to provide item recommendations or predictions based on the opinions of other like-minded users. Collaborative filtering typically suffers three major drawbacks:

1. Cold start problem: The cold start problem appears when new users or items are introduced in the system so that there is no previous information about them to make the predictions;
2. Sparsity problem: In recommendation systems of N users and M items, have a potential number of ratings $N * M$. However the actual number of ratings is

much smaller (not even the 10% in the best cases) than the amount of information available for users and items is small (or sparse).

3. Scalability: It refers to the numbers of operations per number of users/items n performed by the algorithm. Specifically, for an algorithm to be scalable then its time-complexity must grow less than n^2 as n increases. This is critical for the environments in which these systems make recommendations where there are millions of users and items. Thus, a large amount of computation power is often necessary to calculate recommendations and, depending on the algorithm, it would be impossible to make predictions for such a large datasets.

In answer to these challenges, we have developed a network-based recommender system that makes scalable and accurate predictions of individuals' preferences. Our approach is based on the assumption that there are groups of individuals and of items (items, books, etc.), and that the preferences of an individual for an item are determined by their group memberships. Importantly, we allow each individual and each item to belong simultaneously to different groups. The resulting overlapping communities and the predicted preferences can be inferred with a scalable variational maximum likelihood (VML) algorithm based on a variational approximation. Our approach enables us to predict individual preferences in large data sets with tens of millions of observed ratings, obtain considerably more accurate predictions than current approaches available for such large data sets.

4.2 Modeling ratings with a mixed-membership stochastic block model

We have N users and M items, and a bipartite graph $R = \{(u, i)\}$ of links, where link (u, i) is the (observed or unobserved) rating of item i by user u . For each $(u, i) \in R$, the rating r_{ui} belongs to some finite set S (such as $S = \{1, 2, 3, 4, 5\}$ or $S = \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$). Given a subset of $R^O \subset R$ of observed ratings, our goal is to classify the users and the items, and to predict the rating r_{ui} of a new link (u, i) for which the rating is not yet known.

We propose the generative model for this problem. There are K groups of users and L groups of items. For each pair of groups (k, ℓ) , the block model like probability matrices $p_{k\ell}(r)$ (one per each different rating $r \in S$) corresponds to the probability of user u in group k rates an item i in group ℓ with rating r .

We allow both users and items to belong to a mixture of groups. Each user u has a vector $\theta_u \in R^K$, where θ_{uk} denotes the probability with which user u belongs to group k . Similarly, each item i has a vector $\eta_i \in R^L$. Given θ_u and η_i , the probability distribution of the rating r_{ui} is then a combination,

$$\Pr[r_{ui} = r] = \sum_{k, \ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r). \quad (4.1)$$

The normalization constraints over $\{\theta\}$ and $\{\eta\}$ parameters are,

$$\forall u : \sum_{k=1}^K \theta_{uk} = 1, \quad \forall i : \sum_{\ell=1}^L \eta_{i\ell} = 1, \quad (4.2)$$

Also for the rating probability matrices $p_{k\ell}(r)$ are normalized in order that for a given distribution of groups for the user and the item of a given edge, the total probability of having a rating would be one:

$$\forall k, l : \sum_r p_{k\ell}(r) = 1. \quad (4.3)$$

Abbreviating all these parameters as $\{\theta\}, \{\eta\}, \{\mathbf{p}(r)\}$, the likelihood of the model is thus

$$P(R^O | \{\theta\}, \{\eta\}, \{\mathbf{p}(r)\}) = \prod_{(u,i) \in R^O} \sum_{k,\ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui}). \quad (4.4)$$

Indeed, given the probability of a given rating in Eq. (4.1) and a set R^O of observed ratings (the training set), the log-likelihood \mathcal{L} of the model is

$$\mathcal{L} = \log P(R^O | \{\theta\}, \{\eta\}, \{\mathbf{p}(r)\}) = \sum_{(u,i) \in R^O} \log \left(\sum_{k,\ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui}) \right) \quad (4.5)$$

where the first summation runs over the observed ratings. Averaging over the space of all possible mixing memberships $\{\theta_u\}$ and $\{\eta_i\}$ and rating probability matrices $\{\mathbf{p}(r)\}$ (similar to Ref. (Guimerà et al. 2012)) is unfeasible in most practical situations. The alternative that we propose here is to obtain the model parameters that maximize the likelihood using a variational approach, and then use those parameters to estimate unobserved ratings. We apply Jensen's inequality to change the log of a sum into a sum of logs, writing

$$\begin{aligned} \log \sum_{k,\ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui}) &= \log \sum_{k,\ell} \omega_{ui}(k, \ell) \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\omega_{ui}(k, \ell)} \\ &\geq \sum_{k,\ell} \omega_{ui}(k, \ell) \log \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\omega_{ui}(k, \ell)}. \end{aligned} \quad (4.6)$$

where $\omega_{ui}(k, \ell)$ is the probability that a given edge (u, i) from is due to groups k and ℓ respectively. The inequality in Eq. 4.6 holds with equality when

$$\omega_{ui}(k, \ell) = \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\sum_{k',\ell'} \theta_{uk'} \eta_{i\ell'} p_{k'\ell'}(r_{ui})}. \quad (4.7)$$

This gives

$$\begin{aligned}\mathcal{L} &= \log P(R^0 | \{\theta\}, \{\eta\}, \{\mathbf{p}(r)\}, \{\omega\}) \\ &= \sum_{u, i | (u, i) \in R^0} \sum_{k\ell} \omega_{ui}(k\ell) \log \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\omega_{ui}(k\ell)}.\end{aligned}\quad (4.8)$$

We then derive update equations for the parameters $\{\theta\}, \{\eta\}, \{\mathbf{p}(r)\}$ by taken derivatives of the log-likelihood (Eq. 4.8).

If λ_u is the Lagrange multiplier for 4.2, we have that

$$\lambda_u = \frac{\partial \log P}{\partial \theta_{uk}} = \sum_{i | (u, i) \in R^0} \sum_{\ell} \omega_{ui}(k, \ell) \frac{1}{\theta_{uk}}.\quad (4.9)$$

Multiplying both sides by θ_{uk} , summing over k , and applying the normalization condition (4.2) gives

$$\lambda_u = \sum_{i | (u, i) \in R^0} \sum_{k\ell} \omega_{ui}(k, \ell) \frac{1}{\theta_{uk}},\quad (4.10)$$

which substituting in Eq. 4.9 gives,

$$\theta_{uk} = \frac{\sum_{i | (u, i) \in R^0} \sum_{\ell} \omega_{ui}(k, \ell)}{\sum_{i | (u, i) \in R^0} \sum_{k\ell} \omega_{ui}(k, \ell)} = \frac{\sum_{i | (u, i) \in R^0} \sum_{\ell} \omega_{ui}(k, \ell)}{d_u},\quad (4.11)$$

where d_u is the degree of the user u in the network.

Applying a similar procedure for the dependency on $\eta_{i\ell}$,

$$\lambda_i = \frac{\partial \log P}{\partial \eta_{i\ell}} = \sum_{u | (u, i) \in R^0} \sum_k \omega_{ui}(k, \ell) \frac{1}{\eta_{i\ell}}.\quad (4.12)$$

we obtain

$$\eta_{i\ell} = \frac{\sum_{u | (u, i) \in R^0} \sum_k \omega_{ui}(k, \ell)}{\sum_{u | (u, i) \in R^0} \sum_{k\ell} \omega_{ui}(k, \ell)} = \frac{\sum_{i | (u, i) \in R^0} \sum_k \omega_{ui}(k, \ell)}{d_i}.\quad (4.13)$$

Finally, if $\lambda_{k\ell}$ is the Lagrange multiplier for (4.3),

$$\lambda_{k\ell} = \frac{\partial \log P}{\partial p_{k\ell}(r)} = \sum_{(u, i) \in R^0 | r_{ui}=r} \omega_{ui}(k, \ell) \frac{1}{p_{k\ell}(r)}.\quad (4.14)$$

Multiplying both sides by $p_{k\ell}(r)$, summing over r , and applying (4.3) gives

$$p_{k\ell}(r) = \frac{\sum_{(u, i) \in R^0 | r_{ui}=r} \omega_{ui}(k, \ell)}{\sum_r \sum_{(u, i) \in R^0 | r_{ui}=r} \omega_{ui}(k, \ell)}.\quad (4.15)$$

Thus 4.11, 4.13, 4.15 and 4.7 are our update equations. The update equations can be solved by following steps: (i) initialize randomly $\omega_{ui}(k, \ell)$; (ii) update $\{\theta_u\}$, $\{\eta_i\}$, and $\{\mathbf{p}(r)\}$ using Eqs. 4.11, 4.13 and 4.15 with fixed $\omega_{ui}(k, \ell)$; (iii) update $\omega_{ui}(k, \ell)$ with fixed parameters using Eq. 4.7. Alternatively, one can: (i) initialize randomly $\{\theta\}$, $\{\eta\}$, and $\{\mathbf{p}(r)\}$; (ii) update $\omega_{ij}(l, \ell)$ using Eq. 4.7; (iii) compute the new values of $\{\theta\}$, $\{\eta\}$, and $\{\mathbf{p}(r)\}$ using Eqs. 4.11, 4.13 and 4.15. In both cases, it is necessary to iterate (ii) and (iii) until convergence. Note that the number of terms in the sums in Eqs. 4.11-4.7 scales linearly with the number of observed ratings (and with the number of users and items). As the set of observed ratings R^0 is typically very sparse (because only a small fraction of all possible user-item pairs are observed), the calculation of the parameters is feasible even for very large datasets.

In summary, our mixed-membership stochastic block model (MMSBM) approach has a double advantage: (i) it uses a model that is realistic and flexible (see Section 4.6); (ii) the algorithm scales with the number of observed ratings (see Fig. 4.1), and is therefore suitable for very large datasets. Additionally, we do not assume that ratings are linearly spaced in the psychological scale of users, that is, giving an item a rating of 5 instead of 1 does not mean you like it five times as much, which is known not to be true (Ekstrand et al. 2011) as it is also endorsed by our results.

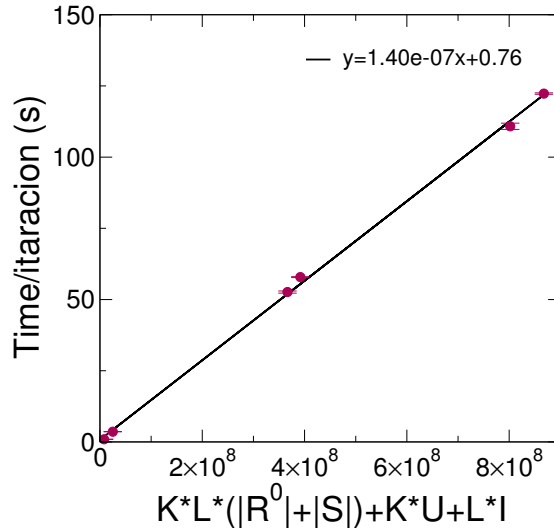


Figure 4.1: **Scalability of the MMSBM algorithm.** Each point represents the average time per iteration in seconds for each of the datasets we use in the study (100K MovieLens, 10M MovieLens, Yahoo! Songs, W-M dating agency, M-W dating agency and Amazon books) each one with different numbers of users U , items I and ratings $|R^0|$ (see Section 4.4). $|S|$ is the number of different ratings values for each recommender systems and K and L are the number of linear fit of the real data, which shows that the computational times per iteration scales linearly with the size of the corpus for the whole range.

4.3 Benchmark algorithms

In this section we present some of the current approaches for recommender systems. We will use them as a benchmark algorithms to validate the results of our MMSBM. Specifically, we consider the following CF-algorithms: the item-item (Sarwar et al. 2001), the singular value decomposition matrix factorization (SVD-MF) (Funk 2006; ?) and stochastic block model recommender algorithm (SBM) (Guimerà et al. 2012). While user-based or item-based collaborative filtering methods (as the item-item) are simple and intuitive, matrix factorization techniques are usually more effective because they allow us to discover the latent features underlying the interactions between users and items. The stochastic block model algorithm is a relatively novel recommendation algorithm that perform a complete statistical treatment following Bayesian approach; it outperforms current algorithms but it is costly in terms of computational time and it can not make prediction on datasets of millions of users (Guimerà et al. 2012). Additionally, we consider a baseline naive model, where the rating of an item by a user is simply the average rating of the item by all users that have rated it before.

Item-item

In a typical collaborative filtering scenario, there is a list of users $U = \{u_1, \dots, u_N\}$ and items $I = \{i_1, \dots, i_M\}$, which the users have rated. The item-item approach assumes that the rating from user u to an item i should be similar to the rating she gives to similar items. Considering the vector $V_i = \{u_1, \dots, u_N\}$ of users that have rated item i , we can obtain the similarity between pairs (i, j) by computing the cosine similarity between V_i and V_j as,

$$\text{sim}(ij) = \cos(i, j) = \frac{V_i \cdot V_j}{\|V_i\|_2 * \|V_j\|_2}, \quad (4.16)$$

or an adjusted version of the similarity (Sarwar et al. 2001). Based on this, only co-rated items has similarity among them. Therefore computing the similarity of the items in the system, and taken the k more similar items, the predicted ratings r_{ui} would be the average of the ratings of users u on the k most similar items,

$$r_{ui} = \frac{\sum_{j \in k \text{ similar items}} (\text{sim}(ij) \cdot r_{uj})}{\sum_{j \in k \text{ similar items}} (|\text{sim}(ij)|)}. \quad (4.17)$$

However if in the k -nearest neighbours there is no item rated by u , the algorithm can not perform a prediction, which may happen for sparse datasets. Also, the algorithm assumes a linear psychological scale on the ratings (that rating 5 is seen as five times better than rating 1), which seems to be a limitation according to our results.

Matrix factorization method based on singular value decomposition (SVD)

One of the widely used recommendation algorithm is the matrix factorization (MF) algorithm (Koren et al. 2009; Paterek 2007). The intuition behind using matrix factorization to solve a recommendation problem is that there should be some latent features

that determine how a user rates an item. Instead of thinking that all ratings on the system are independent, it assumes that there are generalities that guide how users rate on items. In practice this is modelled in MF as free features, shared by users and items, such as the problem is dimensionally reduced. In mathematical terms, MF assumes that the matrix of ratings R (with a number of rows that coincides with the number of users, and a number of columns that coincides with the number of items) can be decomposed into

$$R = P Q, \quad (4.18)$$

where P is a matrix associated with the users and Q is a matrix associated with the items. Each row of the P matrix $\tilde{\theta}_u$ could be seen as a K -dimensional vector with the features values of the user u that describe her, and each column of the Q matrix $\tilde{\eta}_i$ is a K -dimensional vector with the values of the features that describe the item i (with k much smaller than the number of users N and the number of items M). The most efficient method until now to factorize the rating matrix is the singular value decomposition (SVD) (Paterek 2007). This method find the two smaller matrices, which product minimized the error with the original ratings matrix (specifically the means squared error). In addition it uses gradient descent to learn a matrix factorization (trough taken derivatives of the error function over the parameters it tries to infer). The predicted rating would be then,

$$r_{ui} = \sum_k \tilde{\theta}_{uk} \tilde{\eta}_{ik}. \quad (4.19)$$

SVD-MF algorithm is computationally very efficient and performs very good predictions. Also, it has the advantage that it results in intuitive meanings of the resultant matrices. However, features that describe the users and the items are the same; this means that users and items are objects of the same dimension and could be represented geometrical, which could lead to limitation of expressiveness (see Section 4.5.3).

Stochastic block model approach

The stochastic block model (SBM) approach bases its predictions in a family of models (Holland et al. 1983; Nowicki and Snijders 2001) of how social actors establish relationships. In this family of models, social actors are divided into groups and relationships between two actors are established depending solely on the groups to which they belong. The SBM recommender algorithm (Guimerà et al. 2012) assumes that user and items in a bipartite network are connected only depending on their group membership. The ratings in this case are treated as independent labels (ratings 1 and ratings 2 are not considered to be more related than ratings 1 and 5). The algorithm is mathematically sound because it uses a Bayesian approach that deals rigorously with the uncertainty associated with the models that could potentially account for observed users' ratings. Mathematically, the problem is to estimate $p(r_{ui} = r | R^0)$ that the unobserved rating of item i by user u is $r_{ui} = r$ given the observable ratings R^0 ,

$$p(r_{ui} = r | R^0) = \int_M dM p(r_{ui} = r | M) p(M | R^0). \quad (4.20)$$

Where $p(r_{ui} = r|M)$ is the probability that $r_{ui} = r$ if the ratings were actually generated using model M , and $p(M|R^0)$ is the plausibility of model M given the observation. SBM approach averages over the ensemble of all possible generative models M . In practice, the models considered are the family of stochastic block models M_{SBM} , where the of probability that a user rates an item will depends, exclusively, on the groups σ_u and σ_i to which the user and the item belong (one unique group for each user and also one group for each item), that is

$$p(r_{ui} = r) = q_r(\sigma_u, \sigma_i). \quad (4.21)$$

Part of the summation over all possible stochastic block model could be integrated analytically, while the sum over partitions of users and items into groups is estimated by Metropolis-Hastings sampling. Then the prediction for each rating is

$$r_{ui} = \underset{r}{\operatorname{argmax}} p_{SBM}(r_{ui} = r|R^0), \quad (4.22)$$

Importantly, we obtain of the whole probability distribution for each rating (similarly to our MMSBM (Eq. 4.1)). Therefore, we can choose how to make predictions: the most likely rating, the mean, the median, an others. In contrast, recommender systems like MF and item-item give only the most probable rating. However, the SBM approach relies on Markov chain Monte Carlo sampling to make ratings' predictions and therefore does not scale to large datasets. Note that in the SBM the ratings treated independently, without assuming linearity among them.

Naive model

In addition we also perform as a baseline for comparison, a simple model that we call it naive model. In the naive model, the prediction over a ratings r_{ui} would be the average rating received by the item for all the users expect u :

$$r_{ui} = \frac{\sum_{u' \in V_i} r_{u'i}}{|V_i|}, \quad (4.23)$$

where V_i are the users that rate item i , and $|V_i|$ are the number of these users.

4.4 Ratings Data

To validate our model, we perform the predictions on six different data sets: the 100K MovieLens, 10M MovieLens (web site movielens.umn.edu), Yahoo! Songs (Yah), LibimSeti dating agency (Lib) and Amazon books (Ama). We have split the LibimSeti dating agency into two datasets: women rating to men (W-M) dating agency dataset and men to women (M-W) dating agency dataset. There also a total 1% of women rating women and men rating men that we neglected. In table 4.4 we show the characteristics of each dataset.

name	rating scale	#users	#items	#ratings	average fraction of cold start (%)
100K MovieLens	{1, 2, 3, 4, 5}	943	1,682	100,000	0.17%
10M MovieLens	{0.5, 1, ..., 4.5, 5}	71,567	65,133	10,000,000	0.0015%
Yahoo! Songs	{1, 2, 3, 4, 5}	15,400	1,000	311,700	0
M-W dating agency	{1, 2, ..., 9, 10}	220,970	135,359	4,852,455	0.31%
W-M dating agency	{1, 2, ..., 9, 10}	135,359	220,970	10,804,040	0.625%
Amazon book	{1, 2, 3, 4, 5}	73,091	539,145	4,505,893	6.7%

In a cross validation context, the datasets are organized in five different splits, each containing a training set with $\sim 80\%$ ratings and a test set with $\sim 20\%$ ratings. Then the cold start problem is when a user or an item is not in the training but in the test. Since there are 5 training/test, in the table we show the average cold start percentage.

Also the 100K MovieLens dataset provide demographic information of the users: the age in years and the gender. Also for the movies it provide the genre of each movie from a total of 19 labels: Unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War and Western. Each movie could be labelled with more than one genre.

4.5 Results

4.5.1 The MMSBM approach outperforms existing approaches

We test the performance of our algorithm by considering five datasets: the 100K MovieLens dataset (Ekstrand et al. 2011), the 10M MovieLens dataset with (Ekstrand et al. 2011), the Amazon books dataset (Ama ; McAuley et al. 2015), the Yahoo! songs dataset (Yah), and the LibimSeTi dating dataset (Lib ; Brozovsky and Petricek 2007). These datasets are diverse in the kind of items considered, the number of possible ratings, and other factors such as the number and density of observed ratings (see Section 4.4). For purposes of validation, the datasets are organized in five different splits, each containing a training set with $\sim 80\%$ ratings and a test set with $\sim 20\%$ ratings.

We compare our algorithm to benchmark algorithms (see Section 4.3): a baseline naive algorithm that assigns to each test rating r_{ui} the average of the observed ratings for that item i ; item-item (Sarwar et al. 2001; Deshpande and Karypis 2004); and matrix factorization using a stochastic gradient descent learning algorithm (Funk 2006; Paterek 2007). For all these benchmark algorithms we use the implementation in the LensKit package (Ekstrand et al. 2011). Additionally, for the smallest datasets (100K

MovieLens and Yahoo! Songs datasets), we also use the non-scalable SBM approach of Ref. (Guimerà et al. 2012).

For our algorithm, we use a particularly simple version in which the number of groups of users K and the number of groups of items L are both fixed to $K = L = 10$ (to increase the number of groups do not improve the performance significantly but, it increases dramatically the computational time, see Fig. 4.2 for other choices of K and L). Since the iteration of Eqs. (4.11)-(4.7) in general leads to slightly different solutions every run, we obtain the model parameters sampling over 500 times (with different random initialization of the parameters each time) and use an average the predicted probabilities over the 500 runs.

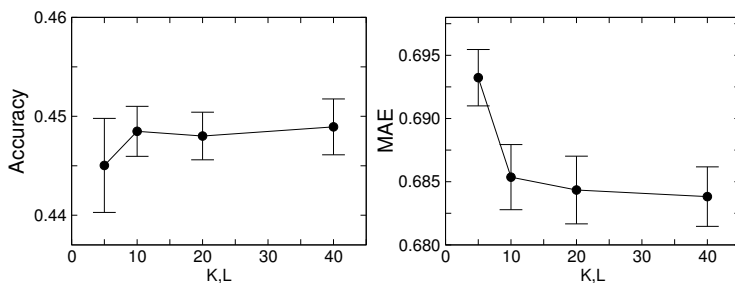


Figure 4.2: **Performance in terms of accuracy and MAE for different groups number for users K and items L .** The results are for the 100K MovieLens dataset. The error bars represent the standard deviation of the sample for the sampling over 500 realizations. The results show that the accuracy and also the MAE performance do not improve significantly after $K = L = 10$.

The performance of the different algorithm is measured in terms of accuracy, that is the fraction of times the predicted ratings are exactly the correct ratings; and we use the mean absolute error (MAE) to measure how close the predictions are. Note that the SMB and our MMSBM give the complete probability distribution of ratings, while the rest of algorithm only give the most likely ratings. Therefore, as the best estimator of the MAE we use: i) the direct prediction from the algorithm for the item-item, matrix factorization and naive model; and ii) we use the median from the complete probability distribution of ratings given by the SBM and our MMSBM, since it performs better than the most likely rating or the mean.

We find that our approach in general outperforms the item-item algorithm and matrix factorization (Fig. 4.3). Indeed, when considering the accuracy, the MMSBM is clearly better than the alternatives in five out of the six examples we consider. The only exception is the Amazon Books database, for which the item-item algorithm performs better than MMSBM. The Amazon dataset presents singularities comparing with the other datasets: first the possibility to rate an article is linked a purchase—users can rate articles only after they buy them; and second the ratings are bias since the users while rating they are seeing at that time the distribution of ratings from other users, which could influence them. The combination of these facts resulted in extremely good ratings for the Amazon dataset comparing with the rest of datasets, since people do not

buy (and then rate) articles with less than three stars over five (the percentage of ratings $r = 1, 2$ is 8.8% while for ratings $r = 4, 5$ the percentage is 80.6%). The item-item algorithm, that performs the predictions based on the average rating of the user to the most similar items, captures better the 'nature' of the Amazon dataset, where the ratings are very correlated. In terms of the MAE, the MMSBM approach is the most accurate in four out of the six datasets (besides the Amazon Books dataset, item-item and matrix factorization produce smaller MAE in the 10M MovieLens dataset).

Importantly, the prediction of the six algorithms are performed same training/test. Some algorithms can not perform predictions in some conditions: for the cold start problem, or because of lack of enough information due to sparsity. We complete the predictions that the algorithms are not able to make in order that all algorithm's predictions are comparable. How the predictions are computed for the cold start problem is detailed explained in the next section 4.5.2. Only the item-item algorithm presents sparsity problem; the algorithm is not able to make a prediction when a user has not rated any of the k -similar items. In these cases we compute the prediction as the average rating that others users give to that item. In any case, both the cold start and the sparsity problems are an small fraction of the whole test (maximum cold start percentage is $< 0.6\%$ for all datasets except for the Amazon that is much higher 6.7%; see Section 4.4 for details), therefore the results do not change significantly with and without these special cases.

Interestingly, our approach produces results that are almost identical to the full non-scalable inference of the SBM for the two examples for which the full inference is feasible. This suggests that the expressiveness of SBMs and MMSBMs is responsible for the high accuracy of these approaches, and that the effect of averaging over user and item partitions as in Ref. (Guimerà et al. 2012) is somewhat equivalent to assuming mixed-membership.

4.5.2 The MMSBM approach provides a principled method to deal with the cold start problem

Cold start problem is a potential problem when the model is not able to draw any inference for users or items about which it has not gather any information. In our recommendation this happens when a user or an item is not in the training set but it is in the test set (in this section we use the same five training/tests as in Fig. 4.3).

Each of the Benchmark algorithms performs differently the cold start problem: i) the item-item do not perform any prediction on the cold start—as if there is no information of the item it can not compute any similarity and make a prediction; ii) matrix factorization and SBM give a prediction of any potential links in the system, included the cold start ratings; and iii) our MMSBM do not include any potential link in the system, but even though it can make predictions on the cold start. Specifically, matrix factorization algorithm includes in the P (for users) and Q (for items) matrices all users and items in the training/test (see MF in Section 4.3), then by SVD predicts a rating for all possible links. The SBM algorithm also is able to make a prediction for the cold

start by giving a group membership to all users and items in the training/test, then the algorithm computes the ratings probabilities for all possible links in the system.

As a base line of the cold start prediction problem, we use the mode of the ratings distributions for each dataset (we check that the mode performs better than the average or the median of the ratings distribution); we call it the most common model. For validation of the cold start, we use the same datasets as in last section (Section 4.4), but note that the Yahoo songs dataset do not have any cold start problem. For all the datasets except the Amazon dataset, the items are always which causes the cold start; in the Amazon dataset there are users and items missing in the training that appear in the test.

Results Fig. 4.4 shown that the MMSBM cold start results bear comparison with the results in 4.3 for the complete training/test, beside the lack of information. Additionally, we observe that our MMSBM outperforms the MF algorithm in all the datasets and in general with wide difference both in terms of accuracy and MAE (except for the 10M MovieLens where the difference is not significant). For the 100K MovieLens dataset, which is the only dataset which we can run the SBM algorithm for the cold start, the accuracy of our MMSBM is similar to the complete probabilistic treatment of the SBM, but our MMSMB presents better MAE. Finally, the most common approach performs very similar to our MMSBM for the W-M dating dataset and the Amazon dataset. From the figure we can read that the $\sim 50\%$ of the W-M dating cold start are women that rate the mode of the ratings distribution to men, also for the Amazon books dataset, the $\sim 56\%$ of the cold start correspond with the mode of the ratings. Remarkably, the MMSBM cold start approach performs very similar for the W-M dating and for the M-W dating, even though the mode of the ratings distribution for the M-W dating dataset only accounts for the $\sim 16\%$ of the cold start, which highlight the robustness of our algorithm.

We conclude our MMSBM approach outperforms the alternatives in most cases for the cold start problem, given robust results for all the datasets. It is done by taking advantage of the information it has from the system and the direct and flexible interpretation of the parameters of the model.

4.5.3 The MMSBM approach highlights the limitations of matrix factorization

As we have discussed earlier (see Section 4.3), a main limitation of the model underlying matrix factorization (when interpreted as a mixture model) is that it assumes that each group of users likes one, and only one, group of items, and totally dislikes the others. The MMSBM relaxes this assumption by introducing the block model like probability matrices, whose element $p_{k\ell}(r)$ is the probability that a user in group k rates an item in group ℓ with a given rating r .

The fact that our approach outperforms matrix factorization suggests that the MMSBM is more expressive than the model underlying matrix factorization; we can check if the introduction of the rating probability matrices is responsible for the improved performance. To this end, we analyze the $\{\mathbf{p}(r)\}$ matrices that maximize the likelihood of the

MMSBM (Fig. 4.5). We observe that, indeed, the observed structure of these matrices is far from the purely diagonal structure implicitly postulated by matrix factorization.

In particular, MMSBMs naturally account for some of the features of real ratings, including: some ratings are much more common than others (for example, $r = 1$ is quite rare whereas $r = 4$ is quite common in Fig. 4.5), but even rare ratings are common among certain groups of items (for example, movies in group $l = 9$ get $r = 1$ quite often) and among certain groups of users (for example, users in group $k = 7$ often give ratings $r = 1$); some groups of users rate most items with the same rating (for example, users in $k = 1$ rate most movies with $r = 5$) and some groups of movies are consistently given the same rating by all users (for example, movies in $l = 3$ are consistently given $r = 5$ by most users); some users agree on rating a group of movies, while disagreeing on others (for example, users in $k = 9$ and $k = 10$ agree at rating with $r = 3$ movies in $l = 8$, but users in $k = 9$ consistently rate movies in $l = 9$ with $r = 1$, whereas users in $k = 10$ consistently rate the same movies with $r = 3$). These observations highlight the limitation in expressiveness of matrix factorization.

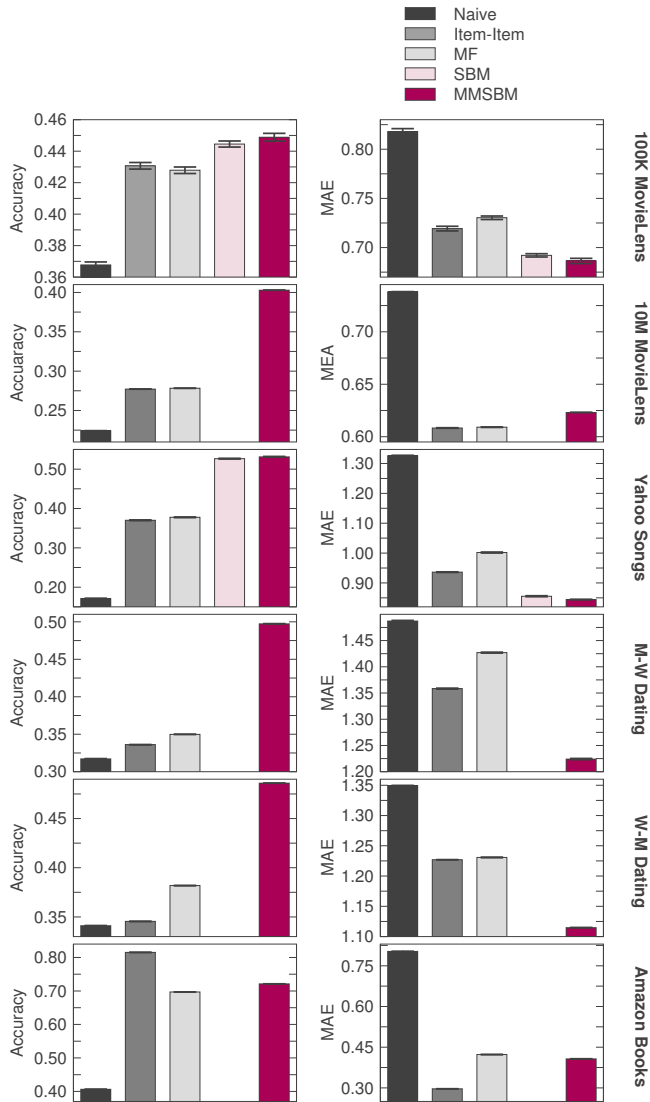


Figure 4.3: **Algorithm comparison for real ratings.** From top to bottom, the 100K MovieLens dataset, 10M MovieLens, Yahoo Song, M-W dating agency, W-M dating agency and Amazon books (details of rating scale, number of users, items and ratings in Section 4.4). The left column graphs are the accuracy for each dataset – that is the fraction of ratings that are exactly predicted by each algorithm. The bars are the average of the 5 training/test from the cross validation and the error bars are the standard deviation of the mean. The right column graphs are the mean absolute error (MAE)– that is the mean absolute deviation of the prediction from the actual rating–, where bars are also the average of the 5 training/test from the cross validation and the error bars are the standard deviation of the mean.

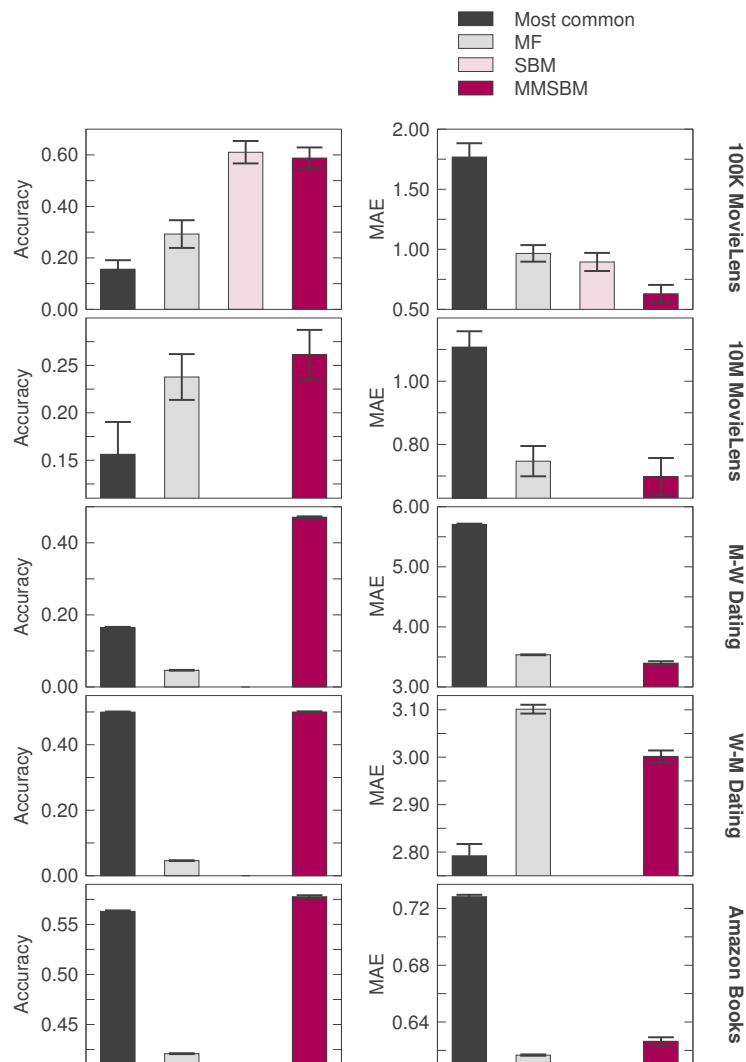


Figure 4.4: **Cold start algorithm comparison for real ratings.** From top to bottom, the 100K MovieLens dataset, the 10M MovieLens dataset, M-W dating agency dataset, W-M dating agency dataset and Amazon book dataset (see Section 4.4 for details of the fraction of cold start for each dataset). Yahoo Songs dataset does not present cold start problem. The left column graphs are the accuracy for each dataset – that is the fraction of ratings that are exactly predicted by each algorithm. The bars are the average of the 5 training-test from the cross validation and the error bars are the standard deviation of the mean. The right column graphs are the mean absolute error (MAE)– that is the mean absolute deviation of the prediction from the actual rating–, where bars are also the average of the 5 training/test from the cross validation and the error bars are the standard deviation of the mean.

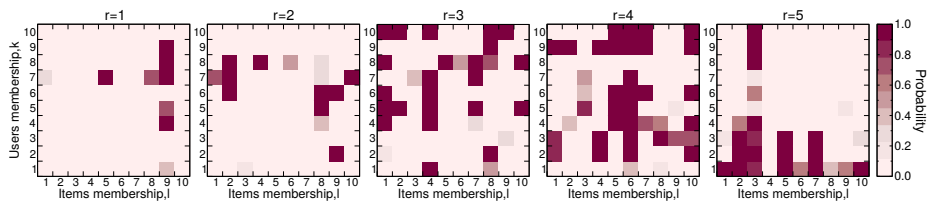


Figure 4.5: **Probability matrices.** The figure shows one of the optimized values for the probability matrices (4.1) for the 100K Movielens dataset. The five matrices correspond with the ratings $r = 1, 2, 3, 4, 5$, and for each of them, the Y axis represents the users groups $k \in K = 10$ and the X axis the items groups $l \in L = 10$. The values of the matrix are the probability that user in group k rate r to items in group l . The matrices are normalized as shown in (4.3). Notice that the probability matrices are not diagonal.

4.6 Groups inferred with the MMSBM reflect trends of users and items

Besides providing accurate predictions of users' ratings, our approach yields, as a by-product, a classification of users and items. An interesting question is whether this classification is related to user and item attributes. Most of the current recommendation algorithms do not provide much information about the users and items as they are more focused in computational performance than in uncovering human behavior, then there is an absence of direct translation of the solutions of those algorithms and the real network. Our MMSM algorithm is based on plausible interpretable assumptions with solid mathematical grounds which made the solution obtained meaningful.

To make meaningful the information from user and items from the algorithms we need outside information. We have information for users and items only from one dataset, the 100K MovieLens dataset (web site movielens.umn.edu). The information provided from the web site contains for the items (movies in this case): each movie is labelled in one or more genre classification; and for the users: for each user there is information about her/his age and her/his gender. From the MMSBM performance, we get the maximum likelihood parameters for users θ_{uk} and items $\eta_{i\ell}$. These parameters could be understood as the probability of users u (movie i) to belong to the group k (group ℓ for the movies), where users and movies could belong to several groups at the same time. We would like to match this information to the outside information given from the system, and going further, to infer possible trends for users and movies.

4.6.1 Inferred trends for movies

The movies present in the MovieLens dataset 100K has outside information about the genre of each movie. There are a total of 19 different labels for the movies' genres: Unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War and Western. Each movie could be labelled with more than one genre, in fact on average there are 1.72 labels per movie. Using this information we study if the movies mixed-membership $\eta_{i\ell}$ are related to the different genres.

For that aim, we analyze if movies in the same genre are more similar among them according to our MMSBM classification, and also if some genres are related between them. First we need to define a similarity measure for the movies. From the MMSBM we get a mixed group membership of each item $\eta_{i\ell}$, which represent the probability of the item i to belong to group ℓ . Based on these parameters, we define the similarity between two movies $\text{sim}_{i,j}$ as the scalar product of the normalized version of η_i , $\tilde{\eta}_i = \eta_i / \|\eta_i\|$ (remember from equation 4.2 it is the sum of the components that equals one, but this is in general different from the norm equal one):

$$\text{sim}_{i,j} = \tilde{\eta}_i \cdot \tilde{\eta}_j = \sum_{\ell} \tilde{\eta}_{i\ell} \cdot \tilde{\eta}_{j\ell}. \quad (4.24)$$

The similarity measure could be understood as the cosine of the angle formed by the normalized group membership vectors $\tilde{\eta}_i$ and $\tilde{\eta}_j$.

In fig.4.6 we show the average similarity for all pairs of movies labelled in the 19 different genres. The results are the average over the 500 realizations of the sampling for similarities between genres. We do not observe any trend. Looking at the diagonal of the heatmap, which represents the average similarity for those pairs that belong to the same genre, the average similarity is not higher. To this point, we conclude that the MMSBM mixed group membership has no relation to these standard labels of genre. In fact, this implies that typically if a user likes (or dislikes) one movie of a genre does not mean that she likes (or dislikes) all movies in that genre.

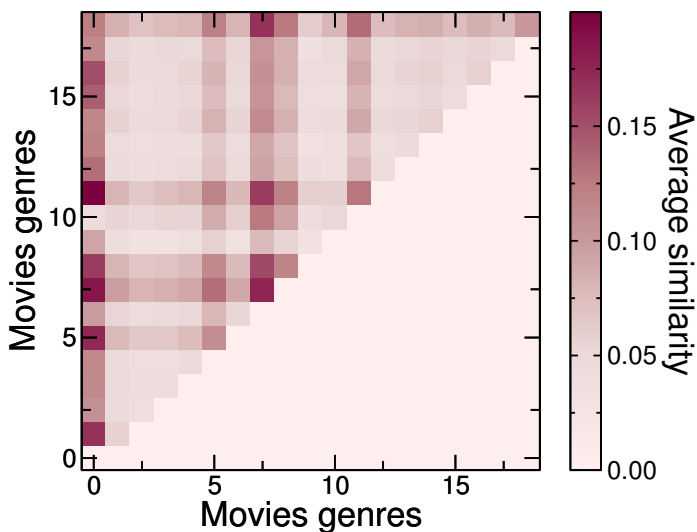


Figure 4.6: **Heatmap of the average similarity between movies' genres.** Darker colors represent higher average similarity between genres. The matrix would be symmetrical, we do not represent the matrix below the diagonal to avoid overload the figure. We do not observe any trend within the genres. Even on the diagonal, that represent the similarity among the same genre, average similarity is not visible higher.

Even though we find that the label of the genre do not match with the mixed group membership, we have checked that the algorithm is performing a 'reasonable' classification. For instance, we find that the popular Star Trek movies 'Star Trek VI: The Undiscovered Country' and 'Star Trek: The Wrath of Khan', seen by hundreds of users, they have a similarity of 0.77. Performing bootstrapping experiments with the similarity values of all pairs of movies, the probability of getting a similarity higher than this by chance is 0.070. Also for the known movies Godfather and Godfather II, also seen by hundreds of users, the similarity between them is 0.84, which has a probability of getting a similarity higher than this by chance of 0.048. According to these, further works should be done to analyze the movie trends. A more plausible approach

should be done to build other categories, not the genre, but related to other reasons that could inspire users to rate movies (such as directors of the movies or the actors and actresses in the cast).

Inferred trends for users

From the 100K MovieLens data sets, the users outside information is the age (in years) and the sex of the users. We want to study if users with similar demographic features present also similar ratings profiles. These patterns could be of interest for the leisure industry and also to have a better knowledge of social trends.

In this case, we are interested in the similarity of users with the same demographic features. We use a similarity measure as the one used for movies. Briefly, from the MMSBM we get the best likelihood group membership of each user θ_{uk} . Based on these parameters, we define the similarity between two users (u, v) as the scalar product of the normalized version of θ_u , $\tilde{\theta}_u = \theta_u / \|\theta_u\|$ as,

$$\text{sim}_{uv} = \tilde{\theta}_u \cdot \tilde{\theta}_v = \sum_k \tilde{\theta}_{uk} \cdot \tilde{\theta}_{vk}. \quad (4.25)$$

First, we analyze how the similarity between users changes with the age. For that aim, we have grouped users for decades according to their age—from 10 to 20, from 20 to 30, etc. (there are only two users under ten and six users over 70, thus they have been added to the nearest group forming finally a total of six age groups). The average similarity in a group is computed as the average over all pairs of users in the same age group.

Note that we obtain the model parameters $\{\eta\}$ and $\{\mathbf{p}(r)\}$ 500 times. Therefore the results in Fig. 4.7 are the average and the deviation over the 500 realizations for similarities in each of the age groups.

In Fig. 4.7 we show a clear trend for the evolution of the inner similarity with the age. The older the users are, the less similar among them. This is a general trend except for those users in the younger group from 10 to 20 years old, where the similarity is still lower than for the next group from 20 to 30 years old, where the average similarity is maximum.

Also of interest is the analysis of the possible differences in similarities depending on the gender. The question to address here is if women are more or less similar among them than men. In Fig. 4.8 we show the average similarities for the same age groups but women with women and men with men. We found that general trend is also valid for both genders, but in different ways. Women starts with a higher similarity from the earliest ages (from 10 to 20), and as the age increase it also increases the differences within women of same age. For men the trend is more similar to the one with all users (Fig. 4.8 in light blue), since most of the users in the ratings systems are men (there are six times more men than women, 54,987 men to 8,247 women). Comparing both, women present more pronounced changes in their ratings profiles with the age (from more similar when they are younger to more different when they become older), while men also are more similar among them when they are younger than when they

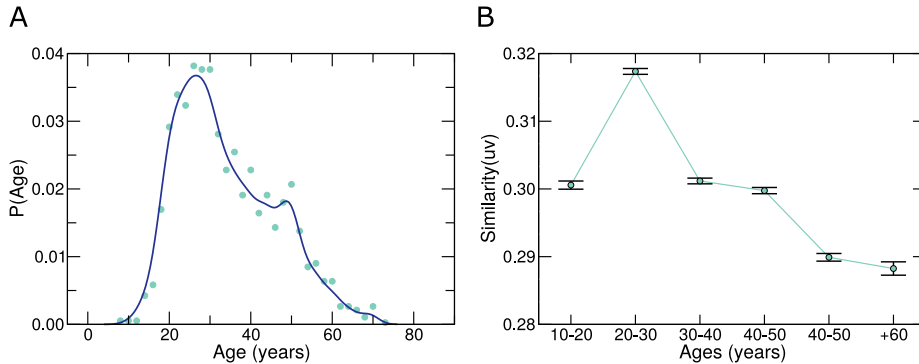


Figure 4.7: **Evolution of the similarity with the age of the users.** (A) The age distribution of the MovieLens dataset. The dots represent the probability distribution of the data and the line is the kernel plot for that distribution. The younger user is 7 years old and the older 73 years old. The peak of the distributions is in the twenties, and from there the number of users decreases. (B) Average similarity of the users grouped by their age buy decades. Notice that as the age of the group increasing, the similarity decrease, while the maximum similarity is achieved in the group of users from 20 to 30 years old.

are older, but less marked. Remarkably, in the group of age from 10 to 20 years old, women's group is more homogeneous than the men's group of the same age. Note that the all users' average similarity (in light blue in Fig. 4.8), is lower than the average similarity among women and among men. This is because the all users values include similarities between women and men, then it should be that the women-men ratings profiles are even more different between them.

4.7 Discussion

We have shown that our mixed-membership stochastic block model in general outperforms the item-item algorithm and matrix factorization, both in terms of accuracy and in mean absolute error, except for the Amazon books dataset. In fact our approach makes predictions that are very similar to the full non-scalable inference of the SBM for the two examples for which the full inference is feasible; even in these cases the MAE of the MMSBM approach is smaller than for the SBM.

In the cold start problem, we found that the predictions of the MMSBM are kept very close to the achieved with the complete training/test beside the lack of information. Moreover, the MMSBM outperforms the MF algorithm in all the dataset both in accuracy and MAE. While the most common model give similar predictions to our MMSBM in some particular datasets, the success of the most common model depends on the ratings distribution and could not be applied systematically, while the MMSBM is more robust.

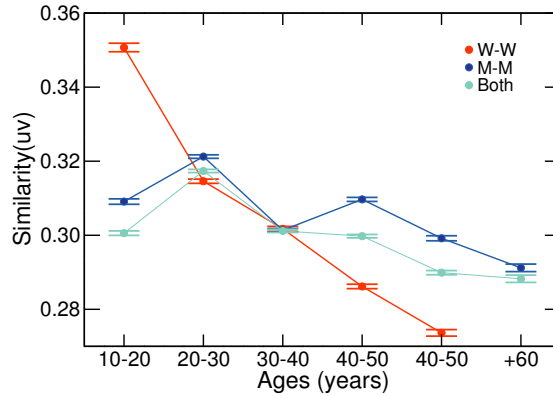


Figure 4.8: **Gender differences in the evolution of the similarity with the age of the users.** Average similarity of the users grouped by their age by decades separated by sex. The red line represents similarities between women, the dark blue represents similarities between men and in light blue there are the similarities among users independently of their sex (same line as in Fig. 4.7B). The similarity is larger between younger groups of users and decrease with the age; but the difference in similarity is more pronounced between women than between men. There are no women with +60 years old in the dataset.

Finally, we found that the parameters of the model allow us to infer trends for users and items given outside information. For the movies, we found that the genre is not a good feature to classify movies by their ratings profile. On the other hand, for the users we found interesting trends in age and gender: i) the younger are the users the more similar rating profiles they have between them; ii) the similarities in the ratings profiles for women among them and men among them follow this general temporal trend; iii) the trend of the similarity with the age is much more pronounced for women than for men.

Temporal inference using mixed-membership tensorial stochastic block models

5.1 Introduction

Most real networks are in constant evolution, and the increasing availability of time-resolved network data sources, e.g., from socio-technical systems and on-line social networks, has brought to the forefront the need to study and understand time varying networks. Temporal dyadic social networks, where the network is defined as the interaction between pairs of users (e.g. an email network, phone call network, internet chats, etc.) are highly unpredictable at the microscopic level, so that in the short term individual ties appear and decay in a highly unpredictable fashion (as shown for the email network in Chapter 3); we also know that at the macroscopic level, beside they display statistical regularities, the social networks are restrained to a "life cycle", that is, to start to grow until it arrives at the maturity and then to decline (Kertesz et al. 2015). However, extracting and characterizing mesoscopic structures in temporal networks would allow us to: i) understand the latent community structure and the temporal activity patterns and ii) predict events in time.

In this chapter we want to model the time evolution of networks to predict events in them. Beside the relevance of the topic, the interest on the temporal inference is fairly recent and therefore there is little work developing rigorous approaches. One of such approaches used Non-negative tensor factorization that perform and uncovered reasonable group formations and time patterns on the scholar schedule (Gauvin L 2014). Also in (Schein et al. 2015) they predict international events using Bayesian Poisson Tensor

Factorization. Another approach use Bayesian inference on layered stochastic block models (Peixoto 2015), in which he is able to find the hidden mesoscopic structure on time-varying networks by identifying the most meaningful time-binning to represent the networks.

Here we propose a new model, the mixed-membership tensorial stochastic block models (MMTSBM). Our model gives a group mixed-membership to users, but importantly, the algorithm also gives mixed groups membership to the time-intervals.

5.2 Modeling temporal inference with a mixed-membership tensorial stochastic block model

Consider a system composed of N users that interact intermittently over a period T . We divide the time period in n_t time-intervals (t_0, \dots, t_{n_T}) . We define an event as an interaction in time t between two users u and v . Then the events are recorded in a tensor of events $A \in \mathbb{R}^{N \times N \times T}$, where each lower dimensional matrix corresponds to the time-interval t , $A^t \in \mathbb{R}^{N \times N}$, at has fulfilled with 1s for events in the time-interval t in the matrix elements (row and column) corresponding to the users that interact and 0s otherwise (A^t is a symmetrical matrix). We try to predict which events between two users (u, v) at time t are more likely.

To solve this problem, we propose the following generative model. There are K groups of users and S groups of time-intervals (note that consecutive time-intervals do not need to be grouped together, we do not impose any structure on the time (or user) groups formation). We allow both users and time-intervals to belong to a mixture of groups. Each user u has a vector $\theta_u \in \mathbb{R}^K$, where θ_{uk} denotes the probability with which user u belongs to group k (idem for the other user v involved in the event, $\theta_{v\ell}$ denotes the probability with which user v belongs to group ℓ , $\ell \in K$). Similarly, each time-interval t has a vector $\tau_s \in \mathbb{R}^S$. For each triad (k, ℓ, s) , there is a probability $p_{k\ell s}$ of an event between a user in group k and user in group ℓ in the time-interval group s . Given θ_u , τ_s and $p_{k\ell s}$ the probability of an event (u, v, t) (meaning an interaction between users u and v in time t) is then the convex combination,

$$p(u \sim v, t) = \sum_{k, \ell, \delta} \theta_{uk} \theta_{v\ell} \tau_{t\delta} p_{k\ell \delta}. \quad (5.1)$$

and the probability of no-event then,

$$p(u \not\sim v, t) = (1 - p(u \sim v, t)). \quad (5.2)$$

The normalization constrains over $\{\theta\}$ and $\{\tau\}$ parameters are,

$$\forall u : \sum_{k=1}^K \theta_{uk} = 1, \quad \forall t : \sum_{s=1}^S \tau_{ts} = 1, \quad (5.3)$$

The likelihood of the model is thus

$$P(A | \{\theta\}, \{\tau\}, \{\mathbf{p}\}) = \prod_{t \in T} \left(\prod_{(u \sim v) \in A^t} p(u \sim v, t) \prod_{(u \not\sim v) \in A^t} (1 - p(u \sim v, t)) \right). \quad (5.4)$$

So that, the log-likelihood \mathcal{L} of the model is

$$\mathcal{L} = \log P(A | \{\theta\}, \{\tau\}, \{\mathbf{p}\}) = \sum_{t \in T} \left(\sum_{(u \sim v) \in A^t} \log \left(\sum_{kls} \theta_{uk} \theta_{vl} \tau_{ts} p_{kls} \right) + \sum_{(u \not\sim v) \in A^t} \log \left(\sum_{kls} (1 - \theta_{uk} \theta_{vl} \tau_{ts} p_{kls}) \right) \right) \quad (5.5)$$

where the first summation runs over the time-intervals, $\sum_{(u \sim v) \in A^t}$ is the sum over all pairs interacting in time-interval t and $\sum_{(u \not\sim v) \in A^t}$ is the sum over all non-interacting pairs in time-interval t . Averaging over the space of all possible mixing weights $\{\theta\}$, $\{\tau\}$ and probabilities $\{\mathbf{p}\}$ (similar to Ref. (Guimerà et al. 2012)) is unfeasible in most practical situations. The alternative that we propose here is to obtain the model parameters that maximize the likelihood using a variational approach (same method as in Chapter 4), and then use those parameters to estimate unobserved events. We apply Jensen's inequality to change the log of a sum into a sum of logs, writing

$$\begin{aligned} \log \sum_{kls} \theta_{uk} \theta_{vl} \tau_{ts} p_{kls} &= \log \sum_{kls} \omega_{uvt}(k, \ell, s) \frac{\theta_{uk} \theta_{vl} \tau_{ts} p_{kls}}{\omega_{uvt}(k, \ell, s)} \\ &\geq \sum_{kls} \omega_{uvt}(k, \ell, s) \log \frac{\theta_{uk} \theta_{vl} \tau_{ts} p_{kls}}{\omega_{uvt}(k, \ell, s)}, \end{aligned} \quad (5.6)$$

and also for the non-events (note that applying normalization Eq. 5.3 $(1 - \theta_{uk} \theta_{vl} \tau_{st} p_{kls}) \equiv \theta_{uk} \theta_{vl} \tau_{st} (1 - p_{kls})$),

$$\begin{aligned} \log \sum_{kls} \theta_{uk} \theta_{vl} \tau_{ts} (1 - p_{kls}) &= \log \sum_{kls} \tilde{\omega}_{uvt}(k, \ell, s) \frac{\theta_{uk} \theta_{vl} \tau_{ts} (1 - p_{kls})}{\tilde{\omega}_{uvt}(k, \ell, s)} \\ &\geq \sum_{kls} \tilde{\omega}_{uvt}(k, \ell, s) \log \frac{\theta_{uk} \theta_{vl} \tau_{ts} (1 - p_{kls})}{\tilde{\omega}_{uvt}(k, \ell, s)}. \end{aligned} \quad (5.7)$$

Where $\omega_{uvt}(k, \ell, s)$ is the probability distribution that a given event (u, v, t) is due to groups k, ℓ and s respectively and $\tilde{\omega}_{uvt}(k, \ell, s)$ is the probability that there is no an event (u, v, t) due to groups k, ℓ and s . These lower bounds hold with equality when

$$\omega_{uvt}(k, \ell, s) = \frac{\theta_{uk} \theta_{vl} \tau_{st} p_{kls}}{\sum_{k' \ell' s'} \theta_{uk'} \theta_{v \ell'} \tau_{st'} p_{k' \ell' s'}}, \quad (5.8)$$

and for the non-events:

$$\tilde{\omega}_{uvt}(k, \ell s) = \frac{\theta_{uk}\theta_{v\ell}\tau_{st}(1 - p_{k\ell s})}{\sum_{k'\ell's'} \theta_{uk'}\theta_{v\ell'}\tau_{st'}(1 - p_{k'\ell's'})}. \quad (5.9)$$

This gives

$$\begin{aligned} \mathcal{L} &= \log P(A|\{\theta\}, \{\tau\}, \{\mathbf{p}\}, \{\omega\}, \{\tilde{\omega}\}) = \\ &= \sum_{t \in T} \sum_{(u \sim v) \in A^t} \sum_{k\ell s} \omega_{uvt}(k\ell s) \log \frac{\theta_{uk}\theta_{v\ell}\tau_{ts}p_{k\ell s}}{\omega_{uvt}(k\ell s)} + \\ &= \sum_{t \in T} \sum_{(u \not\sim v) \in A^t} \sum_{k\ell s} \tilde{\omega}_{uvt}(k\ell s) \log \frac{\theta_{uk}\theta_{v\ell}\tau_{ts}(1 - p_{k\ell s})}{\tilde{\omega}_{uvt}(k\ell s)}. \end{aligned} \quad (5.10)$$

For the maximization, we derive update equations for the parameters $\{\theta\}, \{\tau\}, \{\mathbf{p}\}$ by taken derivatives of the log-likelihood (Eq. 5.10).

If λ_u is the Lagrange multiplier for 5.3,

$$\lambda_u = \frac{\partial \log P}{\partial \theta_{uk}} = \sum_{t \in T} \left(\sum_{v|(u \sim v) \in A^t} \sum_{\ell s} \omega_{uvt}(k, \ell, s) \frac{1}{\theta_{uk}} + \sum_{v|(u \not\sim v) \in A^t} \sum_{\ell s} \tilde{\omega}_{uvt}(k, \ell, s) \frac{1}{\theta_{uk}} \right). \quad (5.11)$$

Multiplying both sides by θ_{uk} , summing over k , and applying (5.3) gives

$$\lambda_u = \sum_{t \in T} \left(\sum_{v|(u \sim v) \in A^t} \sum_{k\ell s} \omega_{uvt}(k, \ell, s) + \sum_{v|(u \not\sim v) \in A^t} \sum_{k\ell s} \tilde{\omega}_{uvt}(k, \ell, s) \right) = T(N-1). \quad (5.12)$$

giving

$$\theta_{uk} = \frac{\sum_{t \in T} \left(\sum_{v|(u \sim v) \in A^t} \sum_{\ell s} \omega_{uvt}(k, \ell, s) + \sum_{v|(u \not\sim v) \in A^t} \sum_{\ell s} \tilde{\omega}_{uvt}(k, \ell, s) \right)}{T(N-1)}, \quad (5.13)$$

Applying a similar procedure for the dependency on τ_t ,

$$\lambda_t = \frac{\partial \log P}{\partial \tau_{ts}} = \sum_{(u \sim v) \in A^t} \sum_{k\ell} \omega_{uvt}(k, \ell, s) \frac{1}{\tau_{ts}} + \sum_{(u \not\sim v) \in A^t} \sum_{k\ell} \tilde{\omega}_{uvt}(k, \ell, s) \frac{1}{\tau_{ts}}. \quad (5.14)$$

Multiplying both sides by τ_{ts} , summing over s , and applying (5.3) gives

$$\lambda_t = \frac{\partial \log P}{\partial \tau_{ts}} = \sum_{(u \sim v) \in A^t} \sum_{k\ell s} \omega_{uvt}(k, \ell, s) + \sum_{(u \not\sim v) \in A^t} \sum_{k\ell s} \tilde{\omega}_{uvt}(k, \ell, s) = \frac{N(N-1)}{2}. \quad (5.15)$$

giving

$$\tau_{st} = \frac{\sum_{(u \sim v) \in A^t} \sum_{k\ell} \omega_{uvt}(k, \ell, s) + \sum_{(u \not\sim v) \in A^t} \sum_{k\ell} \tilde{\omega}_{uvt}(k, \ell, s)}{\frac{N(N-1)}{2}}. \quad (5.16)$$

Finally for $p_{k\ell s}$,

$$\frac{\partial \log P}{\partial p_{k\ell s}} = \sum_{t \in T} \left(\sum_{(u \sim v) \in A^t} \omega_{uvt}(k, \ell, s) \frac{1}{p_{k\ell s}} + \sum_{(u \not\sim v) \in A^t} \tilde{\omega}_{uvt}(k, \ell, s) \frac{1}{(1 - p_{k\ell s})} \right) = 0. \quad (5.17)$$

giving

$$p_{k\ell s} = \frac{\sum_{t \in T} \sum_{(u \sim v) \in A^t} \omega_{uvt}(k, \ell, s)}{\sum_{t \in T} \left(\sum_{(u \sim v) \in A^t} \omega_{uvt}(k, \ell, s) + \sum_{(u \not\sim v) \in A^t} \tilde{\omega}_{uvt}(k, \ell, s) \right)} \quad (5.18)$$

Thus 5.13, 5.16, 5.18, 5.8 and 5.9 are our update equations. The update equations can be solved by following steps: (i) initialize randomly $\omega_{uvt}(k, \ell, s)$ and $\tilde{\omega}_{uvt}(k, \ell, s)$; (ii) update $\{\theta\}$, $\{\tau\}$, and $\{\mathbf{p}\}$ using Eqs. 5.13, 5.16 and 5.18 with fixed $\omega_{uvt}(k, \ell, s)$ and $\tilde{\omega}_{uvt}(k, \ell, s)$; (iii) update $\omega_{uvt}(k, \ell, s)$ and $\tilde{\omega}_{uvt}(k, \ell, s)$ with fixed parameters using Eqs. 5.8 and 5.9. Alternatively, if it is more convenient, we can: (i) initialize randomly $\{\theta\}$, $\{\tau\}$, and $\{\mathbf{p}\}$; (ii) update $\omega_{uvt}(k, \ell, s)$ and $\tilde{\omega}_{uvt}(k, \ell, s)$ using Eqs. 5.8 and 5.9; (iii) compute the new values of $\{\theta\}$, $\{\tau\}$, and $\{\mathbf{p}\}$ using Eqs. 5.13, 5.16 and 5.18. In both cases, it is necessary to iterate (ii) and (iii) until convergence. Note that, both for the parameters update and for the variational distribution functions ω and $\tilde{\omega}$, only the values of the previous step are taken into account. Therefore, each update could easily be computed in parallel, improving considerably the time performance of the algorithm.

5.3 Results

5.3.1 The MMTSBM approach makes good predictions on hidden events

We validate the performance of the MMTSBM on a subset of the emails data. This subset of the email network has 65 users interacting in two months of data (September and October 2010) in times interval of 1 day (61 days in total). We define the network as unweighted and undirected, then an event would be a pair of users sending in any direction one or more emails among them within one day. Therefore, we consider non-events when a pair of users have not send any email in a day. This subset is more dense in events than the whole network, with an average of ~ 25 events per day.

For our algorithm, we use a particularly simple version in which the number of groups of users K and the number of groups for time-intervals S are both fixed to $K = S = 5$, as we consider that the number of users and time-intervals is relatively small (comparing with hundreds to thousands of users of the ratings datasets in chapter 4).

To validate our approach, we hide randomly 20% of the data (events and non-events), and we built 5 training/test sets for a 5-fold cross-validation. The reliability of the possible events in the test is the probability of the event $p(u, v, t)$ as in Eq. 5.2 using the maximum likelihood parameters. To evaluate the performance we use an area under the curve (AUC), that gives the proportion of times the algorithm reliability is correct in assigning a higher reliability to an event than to a non-event. That is done by sorting the reliability of the test's events and non-events. Then the AUC is computed as,

$$AUC = \frac{1}{\#events} \sum_{event_i} \frac{\#non - events\ after(i)}{\#non - events} \quad (5.19)$$

where the sum is only over hidden events and the $\#non - events\ after(i)$ are the number of no-event with lower reliability than the event i on the test. A perfect model will score an AUC of 1, while random guessing will score an AUC of 0.5.

In Fig 5.1 we observe that for each of the training/test the AUC from the cross validation are considerably high, meaning that our MMTSBM is accurate at separating events from non-events. The results are very robust for the 5 training test with very similar results among them, with a total average of $\sim 0.88\%$. That means that from the hidden events and non-events in the test the events has higher reliability than the non-events 88% of the times, even though the data is very sparse, with a proportions fo ~ 1.3 events per 100 non-events. The AUC test confirms the MMTSBM algorithm is able to distinguish accurately events from non-events for this dataset.

5.3.2 Groups inferred with the MMTSBM reflect temporal regularities

Our MMTSBM has the advantage that it is easily interpretable, therefore our approach may be able to shed light into social and psychological processes that determine user behaviors. As said in the previous chapter, relate the parameters of the model with known feature from the reality, we need external data. In the case of the email network, we do not have more information but the actual time resolved correspondence. However, the time dimension is meaningful in itself.

Importantly, contrary to what current approaches do, our algorithm do not assumes that time-intervals to be grouped together should be consecutive. This make our model more expressive and gather more information in a simple manner. Imagine a intermittent periodic email activity between a pair of users: for a model where time-intervals can only be grouped consecutively it would need two groups for each period (one when there is activity and an other when there is not) and it would be necessary to add two groups more for each of the next periods, while in our algorithm could be express only in two time-intervals groups.

We analyze if the time mixed-membership groups unveil inner dynamics in the email correspondence. To do so, we translate the 'comunities' of time-intervals (days in this case) captured by the parameters $\tau_{t,s}$ into the actual days they represent. These parameters could be understood as the probability of time-interval t to belong to the group s , where time-intervals could belong to several groups at the same time. Based

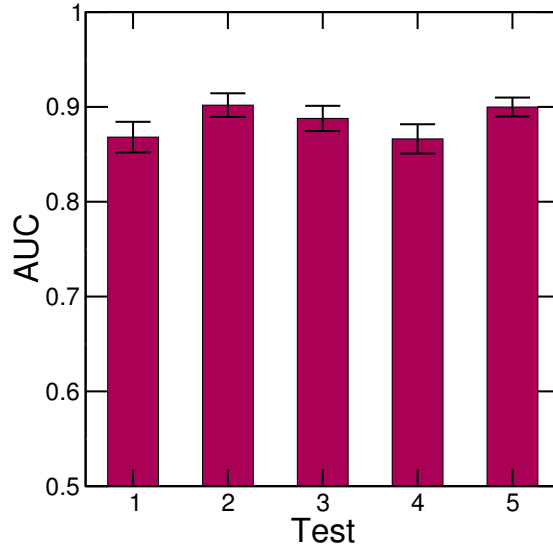


Figure 5.1: **MMSMB performance in AUC test.** The five bars are the AUC performance for each of the 5 training/test of the cross validation. Each Bar is the average AUC for the T=61 time-intervals of the study, and the error is the standard deviation of the mean. A perfect model will score an AUC of 1, while random guessing will score an AUC of around 0.5. The results are very robust for the 5 training/test with a total average over the 5 training/test of 0.885 ± 0.007 .

on these parameters, we define the similarity between two time-intervals $\text{sim}_{t,t'}$ as the scalar product of the normalized version of τ_t , $\tilde{\tau}_t = \tau_t / \|\eta_t\|$ (remember from equation 5.3 it is the sum of the components that equals one, but this is in general different from the norm equal one):

$$\text{sim}_{tt'} = \tilde{\tau}_t \cdot \tilde{\tau}_{t'} = \sum_s \tilde{\tau}_{ts} \cdot \tilde{\tau}'_{t's}. \quad (5.20)$$

The similarity measure is something similar to the cosine of the angle form by the normalized group membership vectors $\tilde{\tau}_t$ and $\tilde{\tau}_{t'}$.

Firts, in Fig. 5.2A we observe how the peaks clearly correspond with a weekly periodicity, then we confirm that the algorithm classification confirms the well-known fact that social communication follow periodicities linked with circadian and weekly cycles of activitiess (Malmgren et al. 2009a; Malmgren et al. 2008; Jo et al. 2012). Then based on the weekly periodicity, in Fig. 5.2B we analyze how similar time-intervals are base on the week days they belong to. We can see two big blocks of higher similarity in the heatmap, one for the working days, and the other for the weekend. Especially Saturday and Sunday are very similar among them. The similarities for the working days in the diagonal (time-intervals of the same day of the week) are not significantly higher than outside the diagonal. Beside the strong peaks found for the similarity with the time difference, it seems that it is not that strong looking at the particular week

days. As a conclusion, we confirm that the time-intervals groups mixed-membership is able to detect unveil temporal regularities of the system. Up to this point, we believe that a longitudinal study for larger periods of time would capture, if there are, long-term temporal regularities of the network.

The detection of unveil temporal regularities is important for prediction. We use the inferred time groups to make predictions, for instance we use what happens one saturday to predict next saturday. Current algorithm do not make use this information.

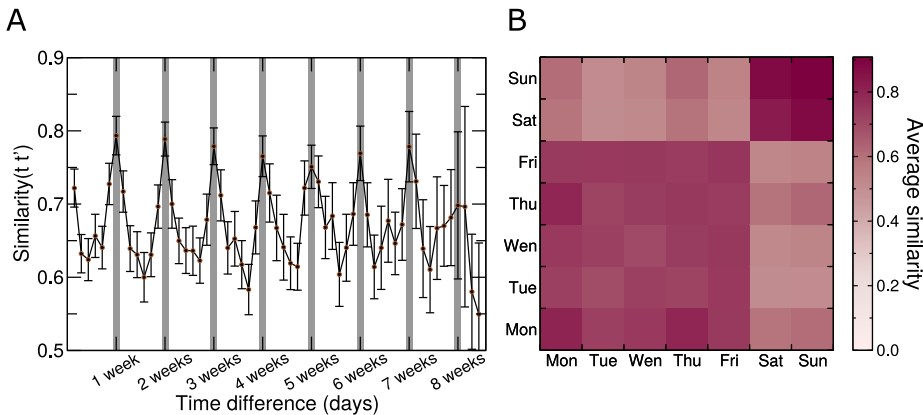


Figure 5.2: **Temporal regularities in the email correspondence.** **A** Average similarity of the time-intervals as a function of the time difference in days. The error bars correspond with the error of the mean. We observe peaks every 7 days, corresponding with a weekly periodicity. **B** Heat map of the average similarity for each pair of time-intervals grouped by the week day they belong. Note that the heatmap is symmetric by definition. There are two blocks of higher similarity corresponding to the working days and the weekend.

5.4 Discussion

We have developed a new algorithm for temporal inference based on mixed-membership tensorial stochastic block model (MMTSBM) with a very good performance in detecting real events. The model assumes mixed group membership on the users that interact and also on the time-intervals, where all of them could belong to different groups, and the interaction between groups is dominated by the block model tensor p_{kls} . From the cross validation, hiding a 20% of the network (events and non-events), we compute the AUC for each of the 5 training test, with high and robust scores. The total average from the cross validation of the AUC is of 0.88 which proves its classification power, even though the data is very sparse.

Moreover, from the MMTSBM parameters' interpretation, we observe temporal regularities on the users' correspondence. We identify weekly periodicity on the time-intervals similarities, and two main blocks in the time-intervals classification, one for

the working days and the other for the weekend, corroborating what previous studies on social communication found about circadian and weekly cycles of users activity in emails networks but also in other communication networks as the call-phone networks (Malmgren et al. 2009a; Malmgren et al. 2008; Jo et al. 2012). This fact highlights the utility of the model parameters to uncover temporal regularities.

Conclusions and perspectives

6.1 Conclusions

The work done in this thesis sheds light on the long-term stability of statistical regularities and patterns in communications networks, and on the predictability of social networks. At the same time, we propose new models for prediction and inference that we validate with real data. The following conclusions can be drawn from the work:

- We have found that the long-term macro-evolution of email networks follows well-defined distributions, characterized by exponentially decaying log-variations of the weight of social ties and of individuals' social strength. These findings imply that fluctuations in connection weights and strengths are considerably larger than one would expect from a process with Gaussian-like fluctuations. Remarkably, together with these statistical regularities, we also observe that individuals have long-lasting social signatures and communication strategies.
- Our results suggest that the existence of correlations is not enough to build a satisfactory predictive model for the logarithmic growth rates. Regarding predictability, we found that a black box method such as Random Forest does not perform better than using the average expected growth for all predictions. The best approach, even with very modest results, is achieved by using the most correlated variables $\omega(t)$ and $s(t)$ for $r_\omega(t+1)$ and $r_s(t+1)$ respectively.

On the other hand, we found that the individual standardized Shannon entropy and strengths could be used to distinguish users among them, even though it

is not enough to uniquely reidentify them. Remarkably, we also found that the combination of different sources of information, in this case the Shannon entropy and the strength, improve significantly the performance in terms of average recall; even when one of them is nearly uninformative.

- We have shown that our mixed-membership stochastic block model in general outperforms the item-item algorithm and matrix factorization, both in terms of accuracy and in mean absolute error. In fact our approach make predictions that are very similar to the full non-scalable inference of the SBM for the two examples for which the algorithm is feasible. Also, as the MMSBM is scalable, it can make predictions on datasets of millions of ratings. For the cold start problem, The MMSBM predictions are kept very close to the achieved with the complete training/test beside the lack of information. Moreover, the MMSBM outperforms the MF algorithm in all the datasets both in accuracy and MAE.

One of the strengths of the MMSBM is that the parameters of the model are interpretable. We have found that the parameters that maximize the likelihood allow us to infer trends for users and items given outside information. For the movies, we found that the genre is not a good feature to classify movies by their ratings profile. On the other hand, for the users we found interesting trends in age and gender: i) the younger are the users the more similar rating profiles they have between them; ii) the similarities in the ratings profiles for women among them and men among them follow this general trend in age; iii) the trend of the similarity with the age is much more pronounced for women than for men.

- We have developed a new algorithm for temporal inference, the mixed-membership tensorial stochastic block model, with a good performance in detecting real events. The model assumes mixed group membership on the users that interact among them and also on the time-intervals, where the interaction between groups is dominated by a block model tensor. We validate the model with email data, by performing AUC experiments with each of the training/test of the cross validation (hiding 20% of events and non-events). We get an average 0.88 AUC score, which proves its classification power even though the data is very sparse.

6.2 Perspectives

Even though the objectives of this thesis has been accomplished, the research developed has opened some interesting questions in the field of social networks. The most relevant in our opinion are:

- They have been reported other human activities that display similar stationary statistical patterns at a macroscopic level (Stanley et al. 1996; Amaral et al. 1997a; Amaral et al. 1997b; Amaral et al. 1998; Plerou et al. 1999). This fact hints the existence of universal mechanisms underlying all these processes (such as, for instance, multiplicative processes (Amaral et al. 1998)). To find

a model of pair communication that reproduces the growth distributions, taking into account the individual signatures and correlation found in this thesis, would be a big step forward in the understanding of human communication.

Additionally, it will be necessary to understand how the individuals' social signatures we observe in the evolution of email networks translate into other types of social networks. All existing evidence suggests that email networks (as well as other techno-social networks such as mobile communication networks (Eagle et al. 2009) and online social networks (Dunbar et al. 2015)) are good proxies for self-reported friendship-based social networks (Wuchty and Uzzi 2011), but more analyses will be necessary to elucidate whether network evolution is also universal. Our findings suggest that may very well be the case.

- We have seen that the combination of the Shannon entropy and the individual strength is a good approach to distinguish users, but it is not enough for a uniqueness reidentification. We believe that the use of external data such as some categorization of the users (these categories should be diverse enough inside the pool of candidates), or inputs of some particular events (like: 'user i was active in the network this particular month/day/hour', or 'user i has in her contact list this other user j '), it would be possible to reidentify uniquely a user with a few additional information.
- One of the strengths of our MMSBM recommender model is that the model itself is interpretable as a mixing of communities. As it has been shown in the thesis that, apart from the communities structure, other trends could be found based on the parameters and some outside information. We believe that there is more information behind the parameters inferred. In the case of the movies, a more complete external information data (as actors, directors or awards of the movies) combined with the model parameters, could unveil unknown trends both for movies and, indirectly, for users.
- It would be interesting to compare our results with other time inference approaches. In addition, we would like to validate the results found by our MMSTBM inference algorithm with different datasets. The datasets should be long enough (several months or years), should have enough time resolution and enough users and events, even though we know that the performance is good enough with sparse data. Also, longer periods of time should be analyzed in order to uncover long-term trends of the temporal networks.

References

- Amazon product data, howpublished = <http://jmcauley.ucsd.edu/data/amazon/links.html>, note = Accessed: 2015-10-06.
- Collaborative filtering dataset - dating agency, howpublished = <http://www.occamslab.com/petricek/data/>, note = Accessed: 2015-10-06.
- Ratings and Classification Data, howpublished = <http://webscope.sandbox.yahoo.com/catalog.php?datatype=r>, note = Accessed: 2015-10-06.
- Albert R., Jeong H., and Barabási A.-L. (1999). Diameter of the World-Wide Web. *Nature* **401**, 130.
- Amaral L. A., Buldyrev S., Havlin S., Leschorn H., Maass P., Salinger M., and Stanley H. E. (1997a). Scaling behavior in economics I: empirical results for company growth. *J. Phys. I France* **7**, 621.
- Amaral L. A. N., Buldyrev S., Havlin S., Leschorn H., Maass P., Salinger M., and Stanley H. E. (1997b). Scaling behavior in economics II: modeling of company growth. *J. Phys. I France* **7**, 635.
- Amaral L. A. N., Buldyrev S. V., Havlin S., Salinger M. A., and Stanley H. E. (1998). Power law scaling for a system of interacting units with complex internal structure. *Phys. Rev. Lett.* **80**, 1385–1388.
- Balcan D., Colizza V., Gonçalves B., Hu H., Ramasco J. J., and Vespignani A. (2009, Dec). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci U S A* **106**(51), 21484–21489.
- Barabási A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211.
- Barabási A. L., and Bonabeau E. (2003). Scale-Free Networks. *Scientific American*, 50—59.
- Barabási A. L., Jeong H., Néda Z., Ravasz E., Schubert A., and Vicsek T. (2002). Evolution of the social network of scientific collaborations. *Physica A* **311**, 590–614.

- Barnes J. (1954). Class and Committees in a Norwegian Island Parish. *Human Relations* (7), 39–58.
- Barrat A., and Barthélemy M.
- Blau P. (1960). A Theory of Social Integration. *The American Journal of Sociology* **65**, 545–556.
- Breiman L. (2001). Random Forests *Mach. Learn.* **45**(1), 5–32.
- Brockmann D., Hufnagel L., and Gaisel T. (2006). The scaling laws of human travel. *Nature* **439**, 462–465.
- Brozovsky L., and Petricek V. (2007). Recommender System for Online Dating Service. In *Proceedings of Conference Znalosti 2007*, Ostrava. VSB.
- Colizza V., Barrat A., Barthélemy M., and Vespignani A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. USA* **103**, 2015–2020.
- de Montjoye Y.-A., Hidalgo C. A., Verleysen M., and Blondel V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* **3**.
- de Montjoye Y.-A., Radaelli L., Singh V. K., and Pentland A. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata *Science* **347**(6221), 536–539.
- Deshpande M., and Karypis G. (2004). Item-Based Top-N Recommendation Algorithms. *ACM TRANSACTIONS ON INFORMATION SYSTEMS* **22**, 143–177.
- Dorfman R. (1979). A Formula for the Gini Coefficient *The Review of Economics and Statistics* **61**(1), 146–49.
- Dunbar R. (1998). The social brain hypothesis. *Evol. Anthr.* **6**(5), 178–190.
- Dunbar R., Arnaboldi V., Conti M., and Passarella A. (2015). The structure of online social networks mirrors those in the offline world *Social Networks* **43**(0), 39 – 47.
- Eagle N., Pentland A., and Lazer D. (2009). Inferring friendship network structure by using mobile phone data *Proc. Natl. Acad. Sci. USA* **106**(36), 15274–15278.
- Easley D., and Kleinberg J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
- Ebel H. D. J., and S. B. (2002). Dynamics of social networks. *Complexity* **8**, 24—27.
- Ekstrand M. D., Ludwig M., Konstan J. A., and Riedl J. T. (2011). Rethinking the recommender research ecosystem: reproducibility, openness, and LensKit. *Proceedings of the fifth ACM Conference on Recommender Systems*, 133–140.
- Funk S. (2006). Netflix update: Try this at home. <http://sifter.org/~simon/journal/20061211.html>, Archived by WebCite at <http://www.webcitation.org/5pVQphxrD>.
- Gautreau A., Barrat A., and Barthélemy M. (2009). Microdynamics in stationary complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8847–8852.

- Gauvin L, Panisson A C. C. (2014). Detecting the Community Structure and Activity Patterns of Temporal Networks: A Non-Negative Tensor Factorization Approach. *PLoS ONE*.
- Girvan M., and Newman M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821–7826.
- Godoy-Lorite A., Guimerà R., and Sales-Pardo M. (2015). Long-term evolution of techno-social networks: Statistical regularities, predictability and stability of social behaviors.
- González M. C., Hidalgo C. A., and Barabási A.-L. (2008). Understanding individual human mobility patterns. *Nature* **453**, 779–782.
- Guimerà R., Danon L., Díaz-Guilera A., Giralt F., and Arenas A. (2003). Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, art. no. 065103.
- Guimerà R., Llorente A., Moro E., and Sales-Pardo M. (2012). Predicting human preferences using the block structure of complex social networks *PLoS ONE* **7**(9), e44620.
- Guimerà R., Mossa S., Turtschi A., and Amaral L. A. N. (2005, May). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. USA* **102**(22), 7794–7799.
- Guimerà R., and Sales-Pardo M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. U. S. A.* **106**(52), 22073–22078.
- Guimerà R., and Sales-Pardo M. (2011). Justice blocks and predictability of U.S. Supreme Court votes *PLoS ONE* **6**(11), e27188.
- Holland P. W., Laskey K. B., and Leinhardt S. (1983). Stochastic blockmodels: First steps *Social networks* **5**(2), 109–137.
- Iribarren J. L., and Moro E. (2009, Jul). Impact of human activity patterns on the dynamics of information diffusion. *Phys Rev Lett* **103**(3), 038702.
- J-P Onnela, Jari Saramäki J. H. G. S. D. L. K. K. J. K. A.-L. B. (2007). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332.
- Jackson M. O., and Watts A. (2002). The Evolution of Social and Economic Networks *Journal of Economic Theory* **106**(2), 265–295.
- Jo H.-H., Karsai M., Kertész J., and Kaski K. (2012). Circadian pattern and burstiness in mobile phone communication *New Journal of Physics* **14**(1), 013055.
- Kertesz J., Lengyel B., Sagvari B., Torok J., and Ruan Z. (2015). The full life cycle of an online social network.
- Koren Y., Bell R., and Volinsky C. (2009, Aug). Matrix Factorization Techniques for Recommender Systems *Computer* **42**(8), 30–37.

- Kossinets G., Kleinberg J., and Watts D. (2008). The structure of information pathways in a social communication network.
- Kossinets G., and Watts D. (2006). Empirical Analysis of an Evolving Social Network. *Science* **311**, 88–90.
- Liljeros F., Edling C. R., Amaral L. A., Stanley H. E., and Åberg Y. (2001). The web of human sexual contacts: Promiscuous individuals are the vulnerable nodes to target in safe-sex campaigns. *Nature* **411**, 907–908.
- Liljeros F., Edling C. R., and Amaral L. A. N. (2003). Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes Infect.* **5**, 189–196.
- Lo A., Chernoff H., Zheng T., and Lo S.-H. (2015). Why significant variables aren't automatically good predictors *Proceedings of the National Academy of Sciences* **112**(45), 13892–13897.
- Lévi-Strauss C. (1947). Les structures élémentaires de la parenté.
- Malmgren R., Stouffer D., Motter A., and Amaral L. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18153–18158.
- Malmgren R. D., Hofman J. M., Amaral L. A. N., and Watts D. J. (2009). Characterizing individual communication patterns. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 607–615.
- Malmgren R. D., Ottino J. M., and Amaral L. A. N. (2010). The role of mentorship in protege performance. *Nature* **465**, 622–626.
- Malmgren R. D., Stouffer D. B., Campanharo A. S. L. O., and Amaral L. A. N. (2009a, September). On Universality in Human Correspondence Activity *Science* **325**(5948), 1696–1700.
- Malmgren R. D., Stouffer D. B., Campanharo A. S. L. O., and Amaral L. A. N. (2009b). On universality in human correspondence activity. *Science* **325**, 1696–1700.
- McAuley J., Pandey R., and Leskovec J. (2015). Inferring Networks of Substitutable and Complementary Products. pp. 785–794.
- McPherson M., Smith-Lovin L., and Cook J. M. (2001). Birds of a Feather: Homophily in Social Networks *Annu. Rev. Sociol.* **27**(1), 415–444.
- Miritello G., Lara R., Cebrian M., and Moro E. (2013). Limited communication capacity unveils strategies for human interaction. *Sci. Rep.* **3**, 1950.
- Miritello G., Moro E., Lara R., Martínez-López R., Belchamber J., Roberts S. G., and Dunbar R. I. (2013). Time as a limited resource: Communication strategy in mobile phone networks. *Soc. Networks* **35**, 89–95.
- Nadel S. (1957). The Theory of Social Structure.

-
- Newman M. E. J. (2001). Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64**, 23102.
- Newman M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.
- Newman M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69**, art. no. 066133.
- Nowak M. A. (2006). Five Rules for the Evolution of Cooperation. *Science* **314**(5805), 1560–1563.
- Nowicki K., and Snijders T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**(455), 1077–1087.
- Oliveira J. G., and Barabási A.-L. (2005). Darwin and Einstein correspondence patterns. *Nature* **437**, 1251.
- Paterek A. (2007). Improving regularized singular value decomposition for collaborative filtering. In *Proc. KDD Cup Workshop at SIGKDD'07, 13th ACM Int. Conf. on Knowledge Discovery and Data Mining*, pp. 39–42.
- Peixoto T. P. (2015, Oct). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Phys. Rev. E* **92**, 042807.
- Plerou V., Amaral L. A. N., Gopikrishnan P., Meyer M., and Stanley H. E. (1999). Similarities between the growth dynamics of university research and of competitive economic activities. *Nature* **400**, 433–437.
- Resnick, P. I. N. S. M. B. P., and Riedl J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *CSCW '94: Conference on Computer Supported Cooperative Work*, 175—186.
- S. H. Strogatz D. J. W. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**(6684), 440–442.
- Saramäki J., Leicht E., López E., Roberts S. G., Reed-Tsochas F., and Dunbar R. I. (2014). Persistence of social signatures in human communication. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 942–947.
- Saramäki J., and Moro E. (2015). From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *Eur. Phys. J. B* **88**(6), 164.
- Sarwar B., Karypis G., Konstan J., and Riedl J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web, WWW '01, New York, NY, USA*, pp. 285–295. ACM.
- Schein A., Paisley J., Blei D. M., and Wallach H. (2015). Bayesian Poisson Tensor Factorization for Inferring Multilateral Relations from Sparse Dyadic Event Counts. pp. 1045–1054.
- Schwarz G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.

- Stanley M. H. R., Amaral L. A. N., Buldyrev S. V., Havlin S., Leschhorn H., Maass P., Salinger M. A., and Stanley H. E. (1996). Scaling behaviour in the growth of companies. *Nature* **379**, 804–806.
- Wellman B. (1988). *Social Structures: A Network Approach*. Cambridge University Press, 39–58.
- Wuchty S., and Uzzi B. (2011). Human communication dynamics in digital footsteps: A study of the agreement between self-reported ties and email networks *PLOS ONE* **6**(11), e26972.
- Yitzhaki S. (1979). Relative Deprivation and the Gini Coefficient. *The Quarterly Journal of Economics* **93**, 321–324.