POLITECNICO DI TORINO

Corso di Laurea in Physics of Complex Systems

Tesi di Laurea Magistrale

Dynamics of Online Social Networks

Cascading phenomena and collapse

Relatore
prof. Andrea Pagnani

Correlatore:
prof. János Kertész

Ludovico Napoli

Anno accademico 2017 – 2018

Summary

Threshold effects in networks trigger global cascades which may generate the failure of such systems. This phenomenon was observed in the case of iWiW, a very popular Hungarian online social network which collapsed due to the cascading abandon of the service by its users, driven by exogenous and endogenous factors. In this research, we analyze the dataset of iWiW and try to characterize some of the dynamical features of the networks, basing our study on the timestamped interactions between users. We first look at the degree distribution of the network and then focus on identifying some of the ego-centered networks. We will then detect the communities in each ego-centered network and analyze the rank correlation between the registration dates and the last login dates of users inside the communities. We find out that the communities have some particular dynamical features but that these are not easily related to their qualitative features (measured with the help of some metadata), pointing out that we need a more refined analysis to reach a conclusion on this particular issue. After that, we still look at the rank correlations for the nodes in some specific paths in the network and show that there is a strong tendency towards anticorrelation. We then develop a criterion according to which a node is considered as part of a departure cascade or not; applying the criterion to all the nodes, we are able to reconstruct the entire cascades history and look at some of their main statistical properties. Lastly, we illustrate the results of the simulations of a simple model that can explain some general dynamical features of such cascading effects on online social networks.

Contents

List of Tables 4

List of Figures 5

1 Introduction 9
- 1.1 Social networks . . . 9
- 1.2 Cascades . . . 11
  - 1.2.1 Information cascades . . . 14
- 1.3 The data . . . 15

2 Ego-centered networks, communities and paths 21
- 2.1 Degree distribution . . . 23
- 2.2 Community detection . . . 23
- 2.3 Rank correlation . . . 25
- 2.4 Overlap . . . 30
- 2.5 Paths . . . 32

3 Cascades 35
- 3.1 The criterion . . . 35
- 3.2 Aggregate results . . . 39
- 3.3 Identifying the cascades . . . 40
- 3.4 Results . . . 41
- 3.5 Discussion . . . 43

4 The model 45
- 4.1 The iWiW case and the model . . . 45
- 4.2 Results . . . 47
- 4.3 Non-interacting mean field model . . . 49
- 4.4 Discussion . . . 53

5 Conclusions 55

Bibliography 57

List of Figures

1.1 Friendship patterns between boys (triangles) and girls (circles) in a class of students in the 1930s (drawing by J. Moreno) [4]. 9
1.2 A portion of the *Facebook* network. 10
1.3 The fraction of individuals joining a riot at equilibrium as a function of the standard deviation of the preferences distribution (Granovetter model) [7]. 12
1.4 Cascade window for the Watts threshold model, for a uniform random graph with homogeneous threshold distribution (all nodes with the same one) [6]. 13
1.5 Different cascade windows by adjusting the parameters p and r in the model of Kertesz et al [14]. 13
1.6 Observing the first k reshares of a cascade, the precision in predicting whether the cascade will double in size increases by observing more of it [12]. 14
1.7 Diffusion tree of a cascade volunteer diffusion protocol on *Facebook* [10]. Early edges are in red while late ones are in blue. 15
1.8 Registered (red) and active (green) users in iWiW [5]. 16
1.9 Google trends of iWiW and Facebook for Hungary [5]. 16
1.10 Cumulative number of inactive users during iWiW activity period (in blue) and quadratic curve $y = 0.227(day)^2$ (in orange). In this plot and in future ones, the day axis starts from the 1st of August, 2007. 17
1.11 The fraction of active users in time: results of the simulations of the threshold model for different parameters (colored curves) compared with the data ("+" dots)[5]. 18
2.1 An example of an ego-centered network in iWiW. 21
2.2 Degree distribution density 22
2.3 Evolution of the average degree for active (green) and registered (red) users [5]. 23
2.4 Community detection with Louvain algorithm of the ego-centered network in 2.1 after removing the ego: different communities are plotted with different colors. 24
2.5 Registration date (x-axis) and last login date (y-axis) plotted together for every user, each in the square relative to the community it belongs to. The colors are the same used in Figure 2.4. 25
2.6 Statistics of the collected rank correlation coefficients inside the communities. Real data are in dark green while the null model (shuffled data) distribution is in light green. 26

2.7 Standard deviation of the correlation distribution for different community sizes. 27
2.8 The variance of the real distribution (blue line) and the variances of the 100 null model distributions (light blue dots). The horizontal axis is just the order of the null models (first value, second value, etc.) 28
2.9 The Shannon entropy of the real distribution (blue line) and the Shannon entropy of the 100 null model distributions (light blue dots). The horizontal axis is just the order of the null models (first value, second value, etc.) . . 28
2.10 The z-score values for the variance of the aggregate distribution (red horizontal line) and of the single degrees distributions (blue dots). 29
2.11 The z-score values for the Shannon entropy of the aggregate distribution (red horizontal line) and of the single degrees distributions (blue dots). . . 30
2.12 Overlap of communities users vs. rank correlations inside the communities. 31
2.13 Rank correlation distribution in cascade paths 32
3.1 P_m shape for different τ 37
3.2 Evolution of the p-value (in semi-logarithmic scale) $W = 200$ during the social network's life. The different curves represent different m_{emp}. The horizontal dashed line is the threshold $y = 0.05$. 38
3.3 In orange, the evolution of the total number of inactive users (the same as the blue curve in Figure 1.10. In blue (dashed), the parabola (orange curve in Figure 1.10). In green, the users with p-value > 0.5 (non cascading). . 39
3.4 Fictitious network: starting node in red; in light blue, nodes that have taken part in the (fictitious) cascade of the starting node; in blue, all other nodes. 40
3.5 The fictitious cascade tree of the starting node (red), obtained from the fictitious graph in Figure 3.4. 40
3.6 The breadth of each layer of ten different cascades. The label in the legend is referred to the I.D. of the starting node (the root of the cascade). 41
3.7 Zoom on the tail of 3.6. 42
3.8 Size of each of the ten analyzed cascades as a function of the last login date of the cascade root. 43
3.9 Graphical representation of the overlap between the analyzed cascades. The node at the tail of an arrow belongs to the cascade of the node at the head of the arrow. 44
4.1 Evolution of a simulation with $p = 0.1$, $\phi = 0.4$ and $k = 0.05$: from left to right, time steps 1, 3, 5 and 7 (the last before the collapse). Red nodes have joined at previous time steps; green nodes are new incomers; blue nodes have left the service at the previous time steps; light blue nodes have just left the service; yellow nodes (very few in this case) are inactive users who join the service again. 46
4.2 ρ dependence on k for different ϕ values. 47
4.3 ρ heat map in the (ϕ, k) plane. 48
4.4 Scaling behavior of the critical value k_c as a function of ϕ. 49
4.5 Evolution of ρ in time during some simulations near k_c for $\phi = 0.4$. . . . 50
4.6 Scaling behavior of the collapsing time. 51

4.7 In green, the phase transition predicted by the non-interacting mean field model. The scattered blue points are the phase transition obtained from the simulations. 52
4.8 Evolution of the registration rate in time (iWiW data) 54

Chapter 1

Introduction

1.1 Social networks

Social networks are one of the most studied types of networks, in which vertices (or *actors*, using sociologists' terminology) are people or groups of people and edges (or *ties*) may represent any pattern of interaction between them [1, 2].

The expression "social networks" commonly refers to online social networking services such as *Facebook* or *Twitter*. But social networks analysis does not necessarily apply to online users' platforms but to any social structure one can think of, in which the connections among actors range from friendships or professional relationships to communication patterns or exchange of money, depending on the specific research goal [1].

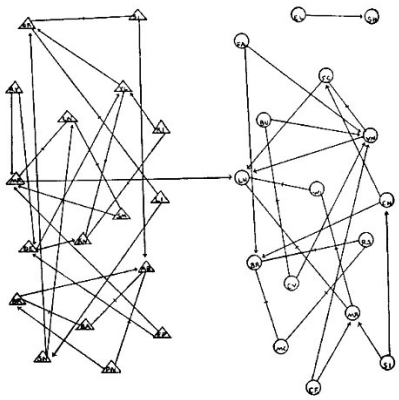


Figure 1.1: Friendship patterns between boys (triangles) and girls (circles) in a class of students in the 1930s (drawing by J. Moreno) [4].

Indeed, the investigation of such systems goes back in the past much further than their recent computer representations. Worth mentioning researches are countless and range from the inspiring work of Jacob Moreno[4] in the 1930s to the famous Milgram experiment in 1961 [3]. In Figure 1.1 we can see what is most probably the first representation of a social network: it has been drawn by Moreno in his book *Who shall survive?*[4] and

it depicts friendship patterns in a class of students, as a result of a real questionnaire (he called this kind of diagrams *sociograms*). It is worth noticing the two distinct communities of boys (triangles) and girls (circles) that emerge. Moreno was interested in the dynamics of social interactions within groups of people and is considered the founder of sociometry.

Nowadays, the great amount of human digital traces that we leave every day is a new incredibly powerful source of information and thanks to the recent developments of computational tools we can address new challenging research questions and find empirical verification of theoretical models. In particular, data of online social networks, of email conversations or of money transactions between people match perfectly well with the abstraction we make by representing the interaction between people as two dots connected by a line. E.g., in Figure 1.2 we can see a beautiful representation of a small portion of the *Facebook* network, where links are the mutual friendships between users (nodes).

Figure 1.2: A portion of the *Facebook* network.

In this research we are going to focus on the phenomenon of cascading failures in online social networks, analyzing the dramatic collapse of iWiW, a popular Hungarian online social network which was active between 2002 and early 2013. We will show the results of some data analysis, trying to get some important properties of the dynamical processes which governed the social network life, and secondly illustrate how the simulations of a simple model can explain some general dynamical features of such cascading effects.

Regarding the data analysis, we will first look at the degree distribution of the entire network and then will focus on identifying some of the ego networks. We will then detect the communities in each ego network with the Louvain algorithm and analyze the rank correlation between the registration dates and the last login dates of users inside the communities. Then, we will look at some metadata (city, age, gender, education level) to see how the characteristics of the users inside a community overlap with the respective ego and if the possible overlap is related to the correlations previously found. After that, we will still look at the rank correlations in some specific paths in the network. Lastly, we will develop a criterion according to which a node is considered as part of a departure cascade or not; applying it to all the nodes, we will be able to reconstruct the cascades and look at some of their statistical properties.

We will first make an introduction on the important phenomenon of cascades in sociology and in network science.

1.2 Cascades

In a social system, actors interact with each other and on many occasions, they are required to take a binary choice. We are continuously exposed to such situations: whether to go to a social event or not, to join a riot, to adopt a new technology, to vote for a certain candidate, to move to another neighborhood or country, to take part in a strike. The decision of actors is not only moved by individual principles and values but is constantly influenced by social pressure. To make a simple example, if a person was attending a boring public lecture but nobody from the audience had left so far, he would hesitate to do it first; but if he noticed that a consistent portion of the attendants had left the room, he would certainly take the option in greater consideration.

The combination of single decisions can generate cascading effects, which are usually very difficult to model and hence to predict. Groups with very similar average preferences may generate very different outcomes, which sometimes appear paradoxical. In the previous example, it could happen that even though the lecture is boring to the majority, everybody hesitates to leave first and stays until the end; on the other side, a few early leavers could convince others to leave and the room will empty very quickly. We can see how the combination of single decisions generates completely different outcomes: even though in both situations the individual perspectives about the lecture were the same, an external observer in the first case would point out how the audience was really captured by the talk (paradoxical).

Even a simple example like this points out how hazardous it is to infer individual dispositions from aggregate outcomes. Since collective behaviors can bring to much more dramatic phenomena than the abandon of a public lecture, sociologists have always been interested in their study. Granovetter in 1978 came out with his famous threshold model, according to which everybody's preferences about an issue are efficiently summarized in a frequency distribution of individual thresholds [7]: in order for an actor to take a certain decision (e.g., to join a riot), the portion of the group that has already chosen that option must exceed his personal threshold. He studied how a given threshold distribution could bring the system to an equilibrium point, but showed how the equilibrium may be extremely unstable by slightly varying the parameters of the distribution. Starting from a Gaussian distribution of individual threshold, in Figure 1.3 the equilibrium fraction of people joining a riot is plotted as a function of the standard deviation of the Gaussian: we notice that there is a critical point where the outcome of the dynamics varies abruptly by slightly changing the preference distribution.

Schelling used a very similar approach in his research on racial segregation in the U.S. and came out with surprising results deriving from collective behavior [8]. In his model, even if agents were inclined to integration (up to a critical point), requiring less than half of their neighbors to be of their same kind, the evolution of the system converges to a segregated configuration, with an average fraction of same kind neighbors much higher than the individual actual requirement. Again, inferring the individual preference or prejudice from the collective outcome would bring to misleading conclusions.

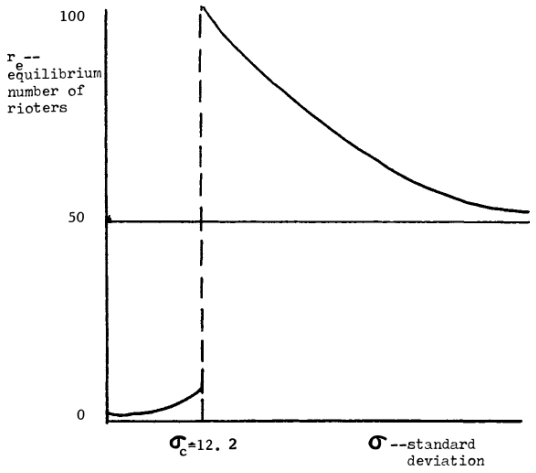


Figure 1.3: The fraction of individuals joining a riot at equilibrium as a function of the standard deviation of the preferences distribution (Granovetter model) [7].

The rise of network science offered to collective behavior and threshold models a new powerful tool, adding to previous studies the underlying network structure which is typical, among others, of social systems. Thanks to the universal nature of networks, it is possible to study and generalize very different cascading phenomena at the same time, although they are generated by quite different mechanisms: from the spreading of information to the financial contagion in networks of banks[16] and the failures in physical infrastructure networks and complex organizations.

Inspired by the previous work of Granovetter and Schelling, Watts implemented an elegant model of cascading behavior on networks, showing that a global cascade (occupying a macroscopic fraction of nodes) can occur from a small initial shock due to the interplay of individual threshold and network structure [6]. In the initial configuration, all nodes start in a state 0 except a small initial seed of nodes in state 1 (adopters). A node with degree k switches to state 1 if the portion of neighbors in state 1 exceeds his individual threshold ϕ. The emergence of a global cascade depends on the degree distribution of the network, the distribution of individual thresholds and the initial seed. The condition for a global cascade is the existence of a percolating component of vulnerable nodes, which are connected to the seed and have thresholds $0 < \phi \leq 1/k$ (needing just one adopting neighbor before exposure). Assuming an Erdős - Rényi random network with average degree z, Watts showed how a phase boundary exists in the (ϕ, z) plane, encompassing a regime where a global cascade can occur (see Figure 1.4).

Kertész et al. generalized Watts model in order to overcome its limitations when compared to real social spreading data [14]. In addition to the individual thresholds, they introduced a rate p of random adoption, such that at each time any node can switch to state 1 with probability p even without the fulfillment of the threshold condition. Another parameter, r, is considered: it is the fraction of "conservative" nodes, who remain blocked in state 0 independently of the dynamics of their neighbors. This model better applies to real systems, where the adoption of a technology in a small city as well as the diffusion of a picture on Facebook can also occur independently of social pressure; at the same time, it

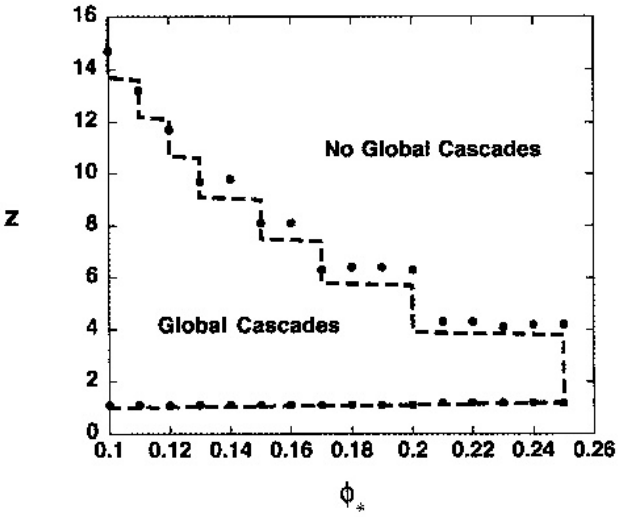


Figure 1.4: Cascade window for the Watts threshold model, for a uniform random graph with homogeneous threshold distribution (all nodes with the same one) [6].

is realistic to consider a small fraction of actors who are reluctant to adopt. The authors showed how the addition of these two parameters modifies the cascading regime area in the (ϕ, z) plane (see Figure 1.5).

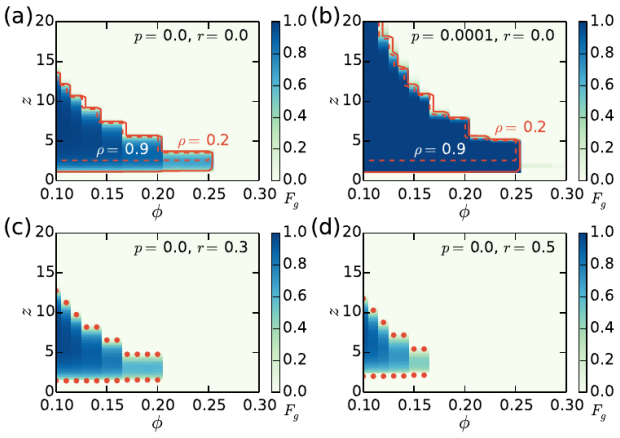


Figure 1.5: Different cascade windows by adjusting the parameters p and r in the model of Kertesz et al [14].

1.2.1 Information cascades

Information cascades are important and deeply investigated dynamical cascading processes in networks. They occur when the diffusion of rumors, photos, disease, memes, or fake news rapidly spreads starting from a small set of nodes in the network, finally encompassing a large fraction of it [17].

Information cascades can be regarded as a manifestation of the robust yet fragile nature of many complex systems: a system may appear stable for long periods with respect to external shocks (robust), then suddenly and apparently inexplicably exhibit a large cascade (fragile)[6]. This kind of cascades is one of the most studied because, behind the interest in the implications of such rare but overwhelming events, there are loads of digital data that can be analyzed.

Since large cascades are very rare (a widespread property that has been observed quantitatively in many systems where information can be shared), they are very hard to predict. It has been shown that e.g. in *Twitter*, although the largest cascades tend to be generated by the most influential users, predictions of which particular user will generate large cascades are rather unreliable [11]. Adamic et al. studied to what extent the future trajectory of a cascade is predictable and which features, if any, are most useful for this prediction task [12], analyzing resharing data on *Facebook*. They found out that the relative growth of a cascade becomes more predictable as more of its reshares are observed (see Figure 1.6) and that, initially, breadth rather than depth is a better indicator of larger cascades; also, they observed that temporal and structural features are key predictors of a cascade size.

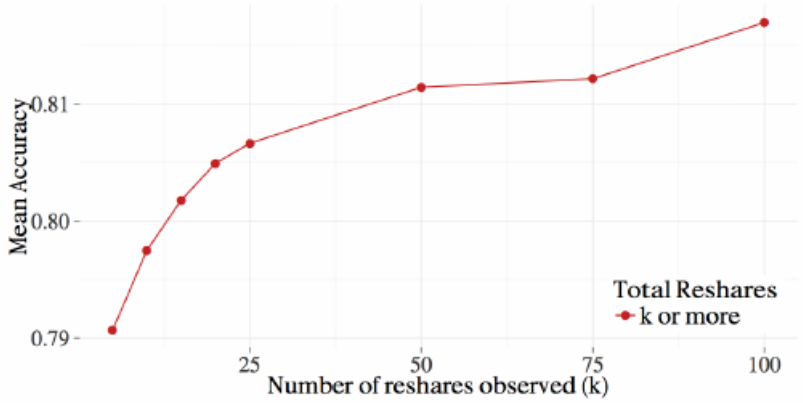


Figure 1.6: Observing the first k reshares of a cascade, the precision in predicting whether the cascade will double in size increases by observing more of it [12].

Besides the fact that most cascades are small and that large cascades are very rare phenomena, the latter also have very different shapes and properties. Dow et al. pointed out that on *Facebook* just a small fraction of photos account for a significant proportion of reshare activity, and they are the ones that generate cascades of non-trivial size and depth [9]. They studied the characteristics of two very large cascades (regarding the reshares of a picture posted by Obama and one by a common user which has become viral) and found

out that they are very different in some of their general properties, such as time evolution, reshare depth distribution, the predictability of subcascade sizes and the demographics of users who propagate them.

Adamic et al. investigated how different diffusion protocols of information resharing on *Facebook* (from tapping a single button in case of a photo resharing to creating and posting a video in the ALS Ice Bucket Challenge) affect the properties of large cascades [10]. In Figure 1.7 we see a representation of one of such cascade with a volunteer diffusion protocols (posting music from an artist whose name matched the letter they were assigned by a friend). They identified two counterbalancing factors (the effort required to participate and the social cost of not participating) that are most influent on the cascade growth and its predictability.

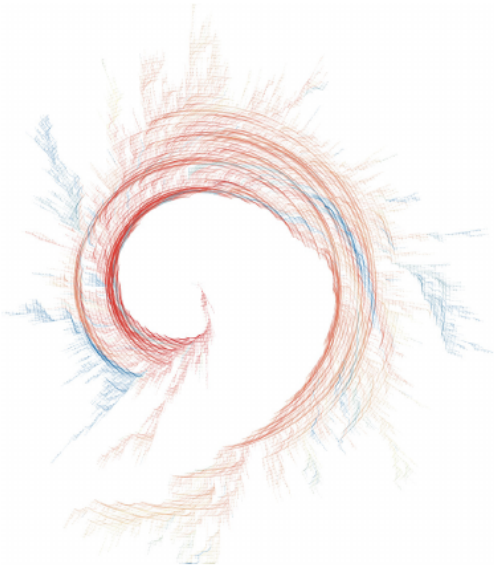


Figure 1.7: Diffusion tree of a cascade volunteer diffusion protocol on *Facebook* [10]. Early edges are in red while late ones are in blue.

1.3 The data

iWiW (*International Who Is Who*) was launched on the 14th of April, 2002, starting as a non-profit project, and shortly became the most known online social network (OSN) in Hungary and even the most visited national website in 2006. The number of users was limited in the first years but started to grow quickly in 2005, probably due to the introduction of new features (e.g. translation into 15 languages, personal advertisements, picture upload, public lists of friends, town-classification, e-mail system, etc.)[15]. In April, 2006 Origo (member of the Hungarian Telecom group) became the owner of the site when the system had 640,000 members with 35 million connections.

The number of registered users continued to rise after that time; it counted for 1.5 million users in December 2006, more than 3.5 million users in October and more than 4 million in December of 2008 [15], in a country with a population of 10 million (worldwide

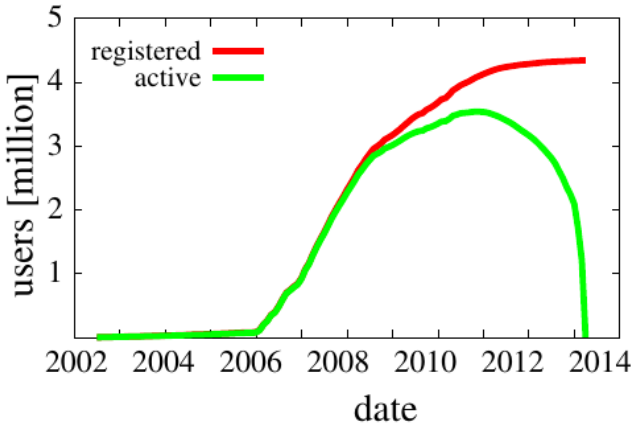


Figure 1.8: Registered (red) and active (green) users in iWiW [5].

13 million native speakers) and, at that time, about 60% Internet penetration. It was certainly the leading social network site of Hungary for years and it is considered as a main driving force behind the speedup of Internet penetration in Hungary [5].

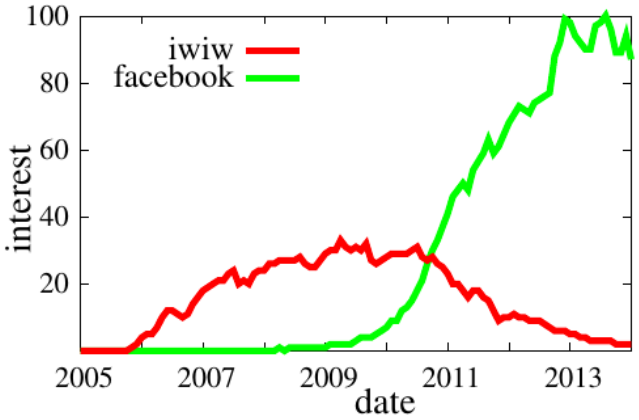


Figure 1.9: Google trends of iWiW and Facebook for Hungary [5].

Until middle 2011 the service was invitation based. Every user, after 30-50 days, got one invitation voucher and new users could register in the service only if they had received a voucher from an already existing member [5]. Later, vouchers were redistributed irregularly, so that users could invite new people without waiting 30-50 days; in the last period after 2012, registration became unconditional (no more vouchers).

This mechanism has two interesting implications: despite the slow down in the growth of the service due to the limited amount of vouchers and the waiting time before getting a new voucher, the site reached a great size and became the most popular OSN in the country; secondly, this limitation makes the data of the early period potentially very

interesting, as we are more confident in assuming that the links represent strong acquaintances (if one has just one shot, it is reasonable to assume that he will invite a close friend).

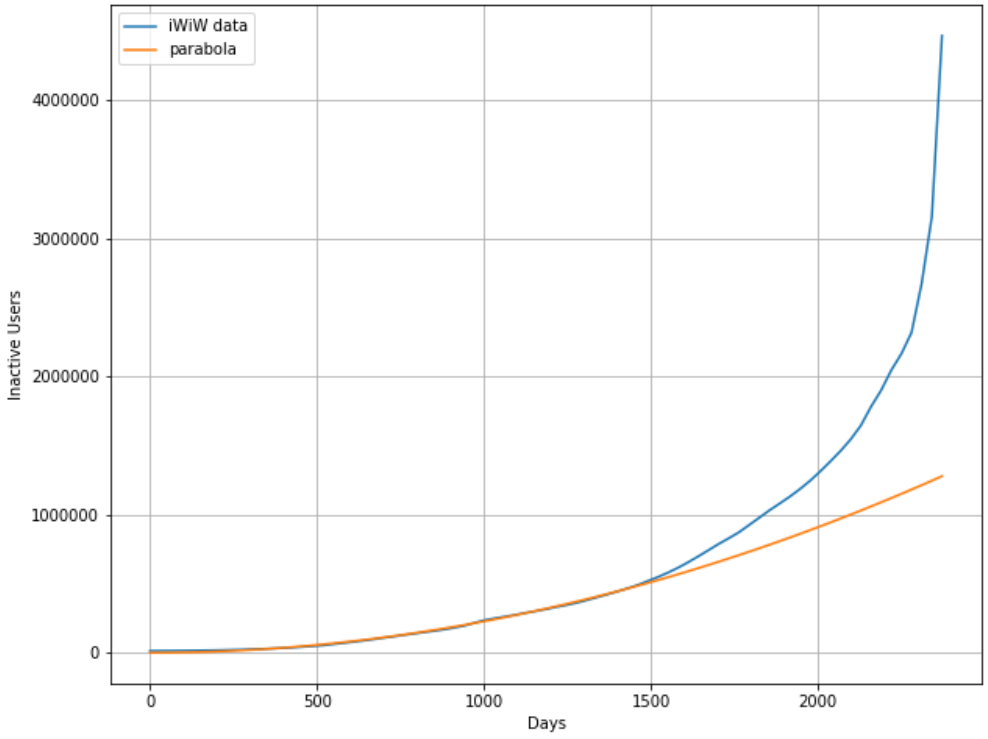


Figure 1.10: Cumulative number of inactive users during iWiW activity period (in blue) and quadratic curve $y = 0.227(day)^2$ (in orange). In this plot and in future ones, the day axis starts from the 1st of August, 2007.

The site has remained widely used even after Facebook became popular (see Figure 1.8 and Figure 1.9) [5]. The story of iWiW came to a sudden and quick end due to various reasons: Facebook became more attractive, especially to young people, after the introduction of games and application; the lack of a usable message filtering system made iWiW a prime target of spammers, which were using mainly compromised accounts; a consistent portion of the Hungarians living abroad also gave a strong push to convert friends to Facebook, which rapidly became the most popular Hungarian OSN. This resulted in a rapid increase in the number of churning users in 2011 and finally led to the collapse in 2012. The site was closed down in June 2014[5].

The blue curve in Figure 1.10 gives us an idea of the sudden breakdown of the service: the number of inactive users (a user is considered inactive after its last login date) started to grow very quickly, especially in the last months where almost all users decided to leave

and the system collapsed.

It is interesting to notice how the number of inactive users grew quadratically until a certain time (the fit with the orange parabola is almost perfect), while after that time it started to grow more rapidly, leading the service to the collapse. The figure suggests that the collapse was driven by two distinct factors, each dominating the dynamics in two distinct periods (before and after the bifurcating point in Figure 1.10). It has been shown that indeed this is the case, and that the churning dynamics was initially driven by random churners, who left the service with a linearly increasing rate, and after the bifurcating point cascading dynamics started to be triggered and gradually started to dominate more and more the churning dynamics, leading to the collapse [5].

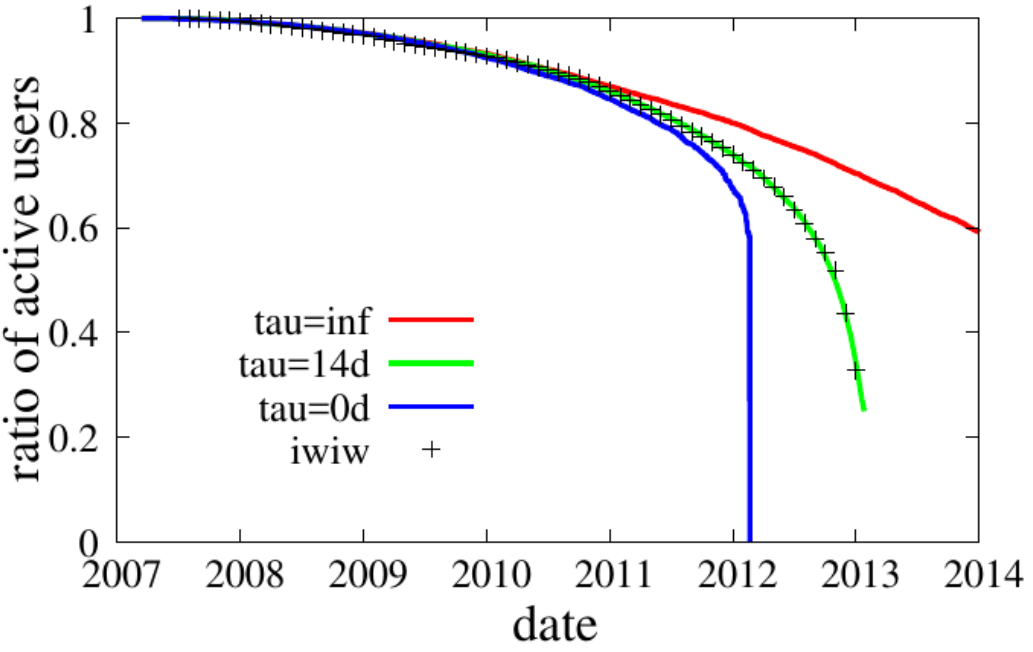


Figure 1.11: The fraction of active users in time: results of the simulations of the threshold model for different parameters (colored curves) compared with the data ("+" dots)[5].

The first process is strictly related to random churners, whose abandon was moved by exogenous factors. The second one is a purely endogenous factor of the network, related to phenomenon according to which a user decides to leave when a certain fraction of neighbors have already left (threshold model). By considering these two combined factors, it has been shown that the simulations of the model fit very well for a certain choice of parameters (the coefficient of linear rate γ and the waiting time τ between the threshold fulfillment and the effective abandon of the user), as shown in Figure 1.11 [5].

The anonymized data used for the following research on the dynamics are: registration and last login dates of each user (when provided) and time-stamped link creation information. Those data are sufficient to reconstruct the whole life of the OSN. Also, in

a small section, we used some metadata of the users when these have been provided: city, age, level of education and gender.

Chapter 2

Ego-centered networks, communities and paths

We start our analysis by looking at some general features of the dynamics of iWiW. Since it is computationally very hard to analyze the whole network (∼4.5 millions of nodes), we start by selecting single nodes and analyzing their ego-centered networks.

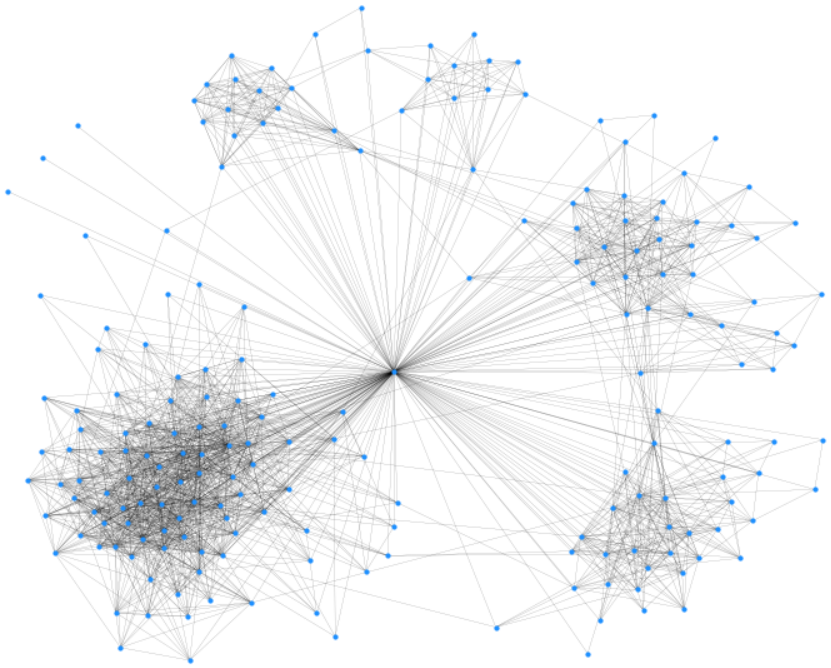


Figure 2.1: An example of an ego-centered network in iWiW.

An ego-centered network is a subgraph of the entire network, whose nodes set is formed by a specific node (the ego) of the network and the group of nodes which are linked to it

(the alters); the edge set is formed by the links between the ego and his alters and the links between the alters.

The latter ones are a natural common characteristic of social networks: two people who are friends with a third person are also likely to be friends. This very small structure forms a triangle and its widespread presence is a well-known property of any social network (and other real systems).

At a larger scale, the effect of these many triangles in social networks is the emergence of communities, which are groups of nodes with many edges between them and comparatively few between them and the nodes of different groups. The emergence of communities is a very particular and well-studied feature of many real networks that is absent, e.g., in random networks models like the Erdős - Rényi or the Barabási - Albert. Their detection is a very important issue in network science and a very hard problem, not satisfactorily solved despite the huge effort of a large interdisciplinary community of scientists.

In Figure 2.1 we can see an example of the ego-centered network of an iWiW user with 200 friends. It is an undirected unweighted network, like all the ones we are going to analyze in this chapter. In this case, it is relatively simple to identify the communities: the clustering structure is quite clear just by looking at the visualization of the graph, without needing the outcome of a sophisticated algorithm.

In this chapter, we would like to understand whether the community structure in the ego-centered networks presents some relevant dynamical features. Also, in the last section, we show how one can find some non-trivial dynamics by going through certain specific paths in the whole network.

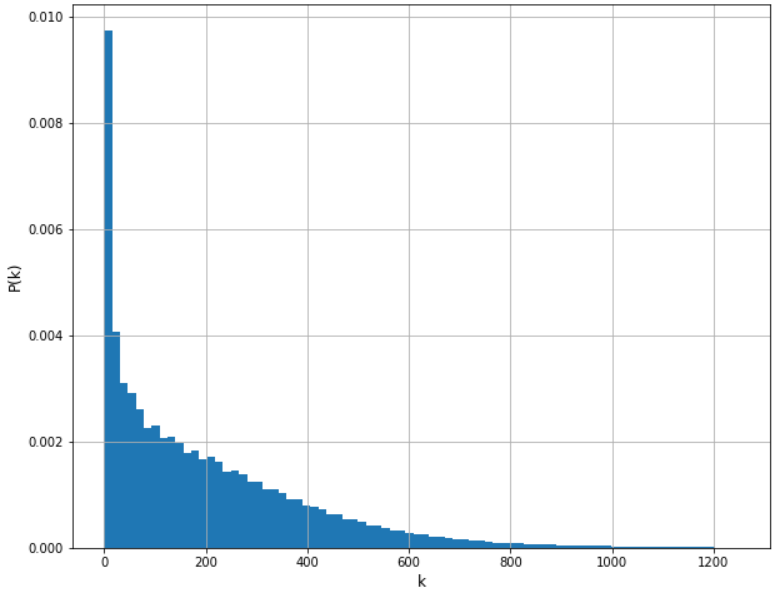


Figure 2.2: Degree distribution density

2.1 Degree distribution

We first look at the general degree distribution of the 4.5 millions vertices of the OSN (regardless of the dynamics) by collecting the respective number of neighbors for each node. The result is plotted in Figure 2.2. We didn't include the non-physical hubs with thousands of acquaintances to make the histogram visible.

As expected, the degree k has a scale-free distribution. The mean degree is 208, but it changes in time. Its evolution is plotted in Figure 2.3. We notice how the average degree of all users (active and inactive) remained constant after 2009 but how the average degree of the active users increased slightly until mid-2011 as shown, indicating that less embedded users left first [5].

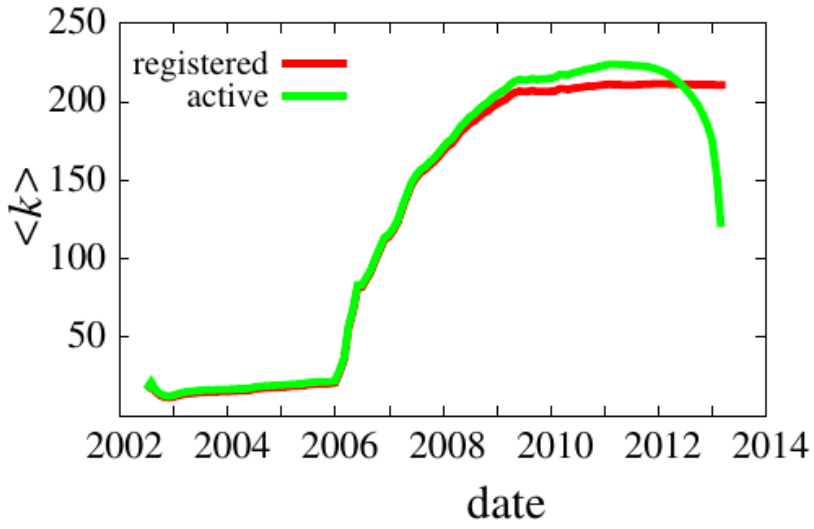


Figure 2.3: Evolution of the average degree for active (green) and registered (red) users [5].

2.2 Community detection

As we mentioned before, community detection is a very hard problem. During the past few years, scientists have developed hundreds of different algorithms. There are many features that qualify the goodness of an algorithm and, indeed, one can be better than another relative to an indicator and worse to another. Depending on the specific problem one is analyzing, he would choose a particular algorithm with good performance in the features of interest.

Many algorithms, like the one we are going to use, are modularity-based, meaning that they identify the communities in a graph by minimizing a function called modularity, whose

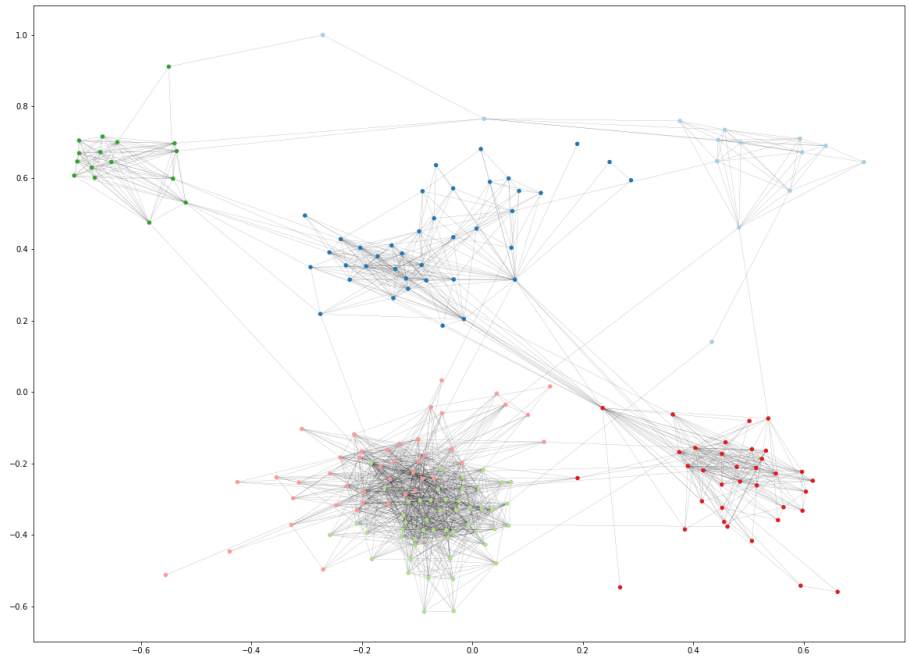


Figure 2.4: Community detection with Louvain algorithm of the ego-centered network in 2.1 after removing the ego: different communities are plotted with different colors.

expression is the following:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, Cj) \tag{2.1}$$

where the sum runs over all pairs of vertices in the graph, A is the adjacency matrix, m the total number of edges and P_{ij} represents the expected number of edges between vertices i and j of the null model. The null model is a random graph with the same feature of the analyzed graph except for the clustering structure, which is not present in random graphs, as mentioned before; it is usually used to quantify the properties of the clustered structure of a given network. C_i and C_j are two subsets of vertices. This class of algorithm compares the number of edges inside the clusters $(C_1, ..., C_N)$ with the expected one in a random graph. The algorithm runs through all the possible configurations of clusters and maximizing this difference (and hence Q) it obtains the desired configuration[18].

The algorithm we are going to use is called Louvain algorithm; it is one of the most used and it works as follows. It is generally applied to weighted networks, but, of course, can be applied also to our unweighted ego-centered networks (an unweighted network can be seen as a weighted network with all the weights equal to 1). Initially, all vertices of the graph

are put in different communities (one community per node). The first step consists of a sequential sweep over all vertices, meaning the following: given a vertex i, one computes the gain in weighted modularity (equation 2.1) coming from putting i in the community of one of its neighbors j; at the end on picks the community of the neighbor that yields the largest increase of Q (as long as it is positive). At the end of the sweep, one obtains the first level partition. In the second step considers the graph where communities are replaced by supervertices and two supervertices are connected if there is at least an edge between vertices of the corresponding communities. The same sequential sweep is applied to the supervertices, obtaining a second level partition (in this case, the weight of the edge between the supervertices is the sum of the weights of the edges between the represented communities at the lower level). The two steps of the algorithm are then repeated, yielding new hierarchical levels and supergraphs, until a single supervertice remains at the last level [18].

In Figure 2.4 we plotted the result of the Louvain community detection algorithm applied to the ego-centered network in Figure 2.1, after removing the ego (since it should be included in every community due to its full connectivity).

2.3 Rank correlation

As a first indicator of the role of communities in the dynamics of the ego networks, we compute the Spearman correlation coefficient (or rank correlation) between the registration dates and the last login dates of users inside the communities, for each community of each selected ego network. To have a heterogeneous sample, we collect 200 ego networks for each degree from $k = 50$ to $k = 400$ with jumps of 50 (hence for $k = 50$,100,150, ...,400).

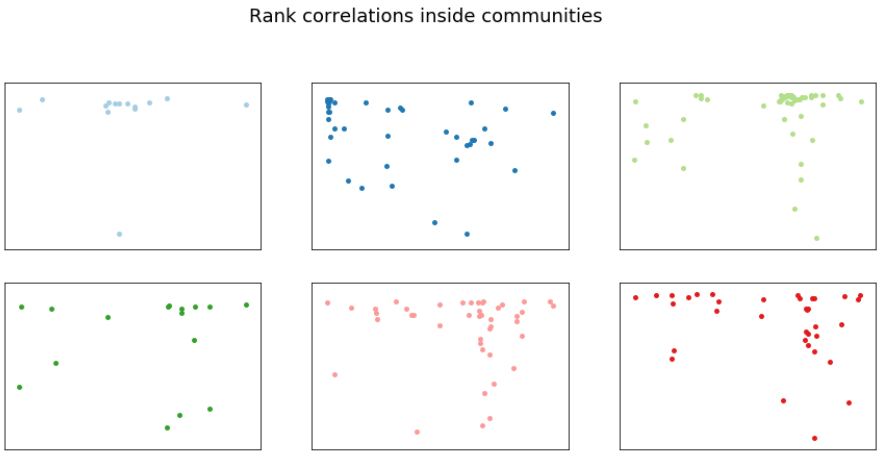


Figure 2.5: Registration date (x-axis) and last login date (y-axis) plotted together for every user, each in the square relative to the community it belongs to. The colors are the same used in Figure 2.4.

The Spearman correlation just takes into account the order (rank) in which two sets of variables are placed, regardless of their value (unlike the Pearson correlation). By

definition, the Spearman correlation is the Pearson correlation for the ranked variables. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks rg_{X_i} and rg_{Y_i} and the rank correlation coefficient r_s is computed from:

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

where ρ denotes the usual Pearson correlation coefficient but applied to ranked variables, as mentioned before; $cov(rg_X, rg_Y)$ is the covariance of the ranked variables; σ_{rg_X} and σ_{rg_X} are the standard deviations of the ranked variables.

In Figure 2.5 we plotted the registration date (horizontal axis) against last login date (vertical axis) for every user inside the communities of the ego network in Figure 2.4. The rank correlation is supposed to measure whether there is an order in these two quantities inside the communities; we are interested in measuring whether this is on average more significant than the expected one (obtained from a null model).

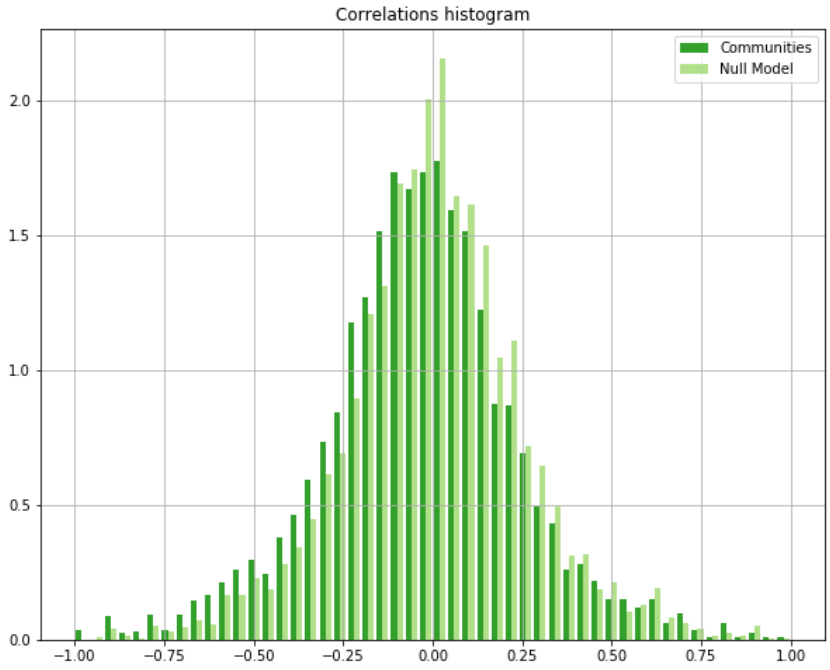


Figure 2.6: Statistics of the collected rank correlation coefficients inside the communities. Real data are in dark green while the null model (shuffled data) distribution is in light green.

Our aim is to collect all these coefficients, to make a statistics and to compare it with the statistics of a null model. Hence, we first collect the coefficients of all the communities of all the 1600 ego-centered networks of our sample and we make the histogram of the

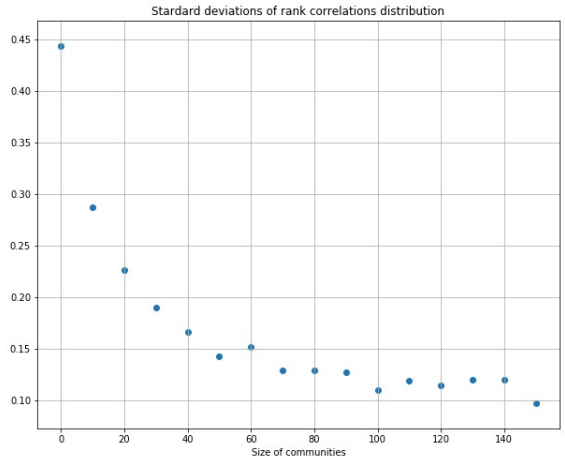


Figure 2.7: Standard deviation of the correlation distribution for different community sizes.

coefficients. Then, for the null model, we shuffle at random the last login date values before computing the rank correlation, so that the order is completely random. If the community structure has some role on the dynamics, we expect the real distribution to be wider than the one relative to the null model (which for a large sample should be peaked around zero).

The comparison of the two distribution is plotted in Figure 2.6. The largest (in absolute value) correlations come from small communities, as we can see in Figure 2.7: we collected the coefficients by different community size and analyzed the different distributions, finding out that the small communities distributions are wider while the large communities ones are more peaked around zero.

We can observe how the null distribution in Figure 2.6 looks slightly more peaked around zero (as expected), while the real one is wider and a bit shifted towards anticorrelation values. The latter characteristic is something we could expect. In fact, inside a community people are supposed to have closer relations, on average, and for a strong community, we expect the following dynamics: I invite my best friend to join the service, my best friend invites his best friend and so on, until the community emerges; then, I leave the service because my best friend has already left because his best friend had already left and so on.

The two chained dynamics (entry and abandon) go in opposite directions, hence in the previous ideal case the rank correlation between registration and last login date would be $r_s = -1$. Of course, this is very simplified, but nonetheless, we expect a tendency towards anticorrelation and the slight shift of the empirical distribution towards negative values seems to confirm our assumptions; nevertheless, this is not so evident and we probably need more accurate analysis in this sense.

We would like to quantify how much wider is the real distribution compared to the random one (null model). To do that, we consider two relevant quantities related to the distributions: the variance and the Shannon entropy. We consider them as random variables and compute their z-score with respect to the distribution of the null model values

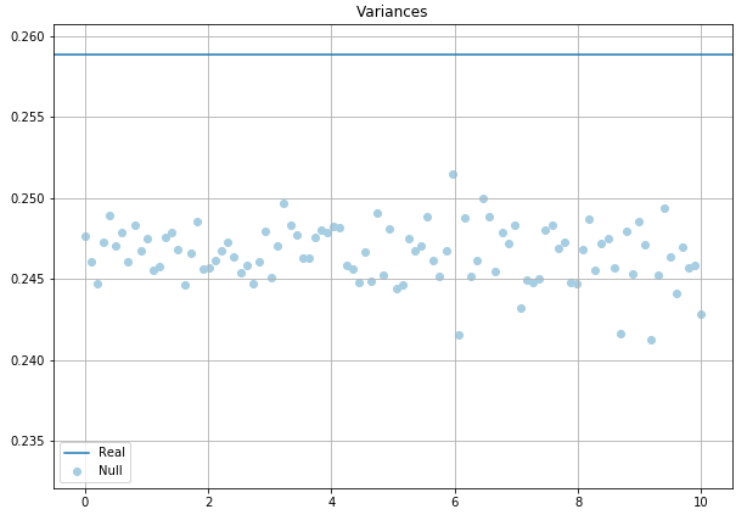


Figure 2.8: The variance of the real distribution (blue line) and the variances of the 100 null model distributions (light blue dots). The horizontal axis is just the order of the null models (first value, second value, etc.)

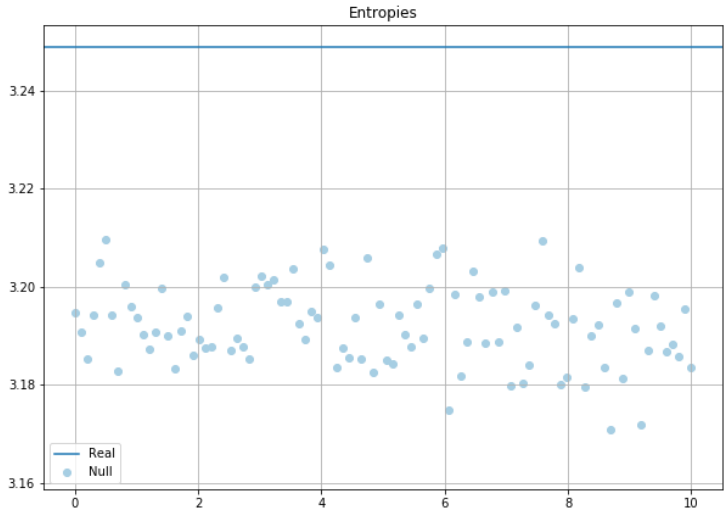


Figure 2.9: The Shannon entropy of the real distribution (blue line) and the Shannon entropy of the 100 null model distributions (light blue dots). The horizontal axis is just the order of the null models (first value, second value, etc.)

of variance and entropy obtained from many null model outcomes. The z-score is defined as:

$$z = \frac{x - \mu}{\sigma}$$

where x is the observed quantity (the variance or the Shannon entropy of the real distribution), μ and σ are respectively the expected value and the standard deviation obtained by the many null model outcomes: by performing the shuffle of the last login dates many times, one obtains different null model distributions; by collecting the variances and the entropies of each of them, we have a distribution of variance and entropy values, from which we can compute the mean value and the associated standard deviation for the calculation of the z-score.

We collect the variances and the Shannon entropies of 100 null model distributions and then compute the mean and the standard deviation of these 100 values, which will correspond to μ and σ.

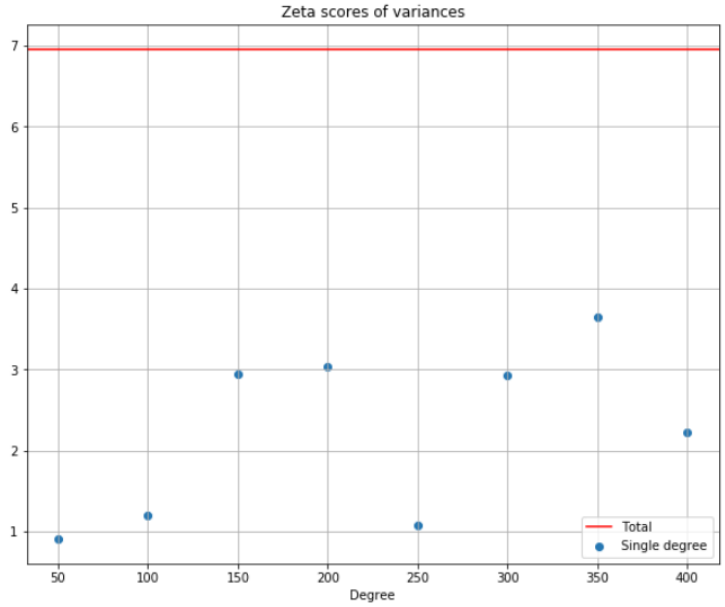


Figure 2.10: The z-score values for the variance of the aggregate distribution (red horizontal line) and of the single degrees distributions (blue dots).

The results are plotted in Figure 2.8 for the variance and in Figure 2.9 for the entropy. We observe how effectively in both cases the real value (the horizontal blue line) is on average consistently higher than the outcomes of the null model distributions (light blue dots), indicating that the dynamics inside communities is not random. The two quantities (variance and entropy) show similar outcomes (besides a slightly higher concentration of the null model values in the variance plot): this is not surprising, as the variance and entropy are two ways of measuring the property we were interested in.

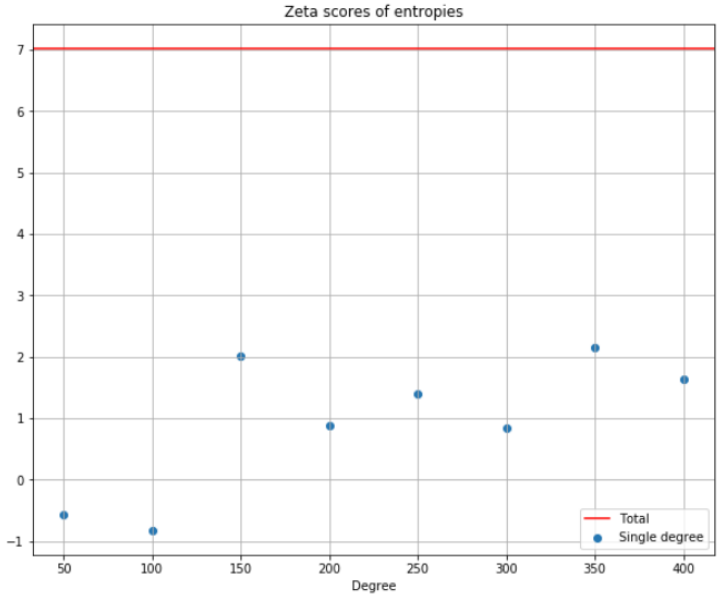


Figure 2.11: The z-score values for the Shannon entropy of the aggregate distribution (red horizontal line) and of the single degrees distributions (blue dots).

We plot the results for the z-score in Figure 2.10 and Figure 2.11, adding the results for the single degree distributions, i.e. the collection of the coefficients of egos with the same degree (50,100,...), to compare them with the aggregate outcome. The horizontal axis is the degree and the dots are relative to the distribution of egos with the same degree, while the red line is relative to the aggregate distribution (see Figure 2.6). We notice how the z-score is very high both for the variance and for the entropy ($\sim$7) and how the aggregate value is always higher than the single degree values.

We have found that the order of registration and last login dates in communities is not random, which is a preliminary but important finding of the dynamics inside the communities of ego-centered networks.

2.4 Overlap

We have seen that the order in which users register and abandon the service is not random inside the community. We would like to see if the rank correlation is somehow related to some general features of the users inside the community, compared to the features of the ego.

To understand if the dynamics of the communities is related to their features, we analyze some metadata: city, age, education level and gender. Unfortunately, not every user has provided all these data and one could address these missing data with some data mining technique; however, this is a preliminary exploration and if there is some sort of effect it will be evident even from the available data.

We introduce a measure that quantifies the similarity between the community and the ego: the overlap. We first look at the metadata of the ego and then focus on the users of

communities. The overlap of a single user related to the ego is just the number of features that this user has in common with the ego (e.g. if he lives in the same city, if he was born no more than two years before or after the ego, ect.), divided by the total number of possible comparisons, to obtain a fraction. Besides the previously mentioned metadata, among the features we analyze to compute the overlap we take a property that is also an indicator of the similarity between two nodes: the number of neighbors they share (friends in common). Once we have the overlap value of all the users in the community, we average all these values to obtain a single overlap between the community and the ego.

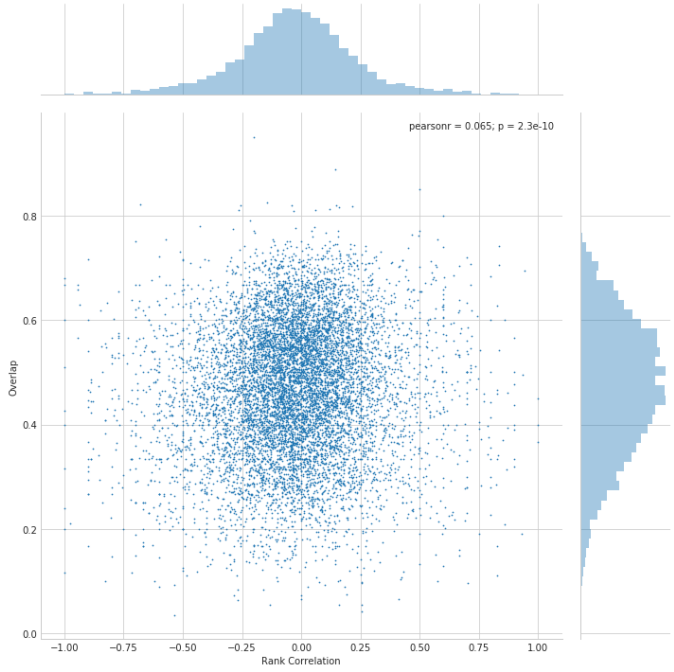


Figure 2.12: Overlap of communities users vs. rank correlations inside the communities.

Our aim is to see whether the qualitative similarity between the community and the ego is also related to the order in which the nodes of the community sign in and log out. We compute the overlap for all the communities whose correlation coefficient has been already collected (see the previous section).

The result is in Figure 2.12, where we plotted the overlap vs. the rank correlation for every community of every ego. As we can see, there is no clear effect but just a big cloud of points. Therefore, we can not conclude anything about the relation between the order of the community and the similarity with the ego.

We wanted to see if the most overlapping communities (which are likely the ones which have more influence on the ego) had a particular dynamical order inside. The problem with this approach is that we didn't filter the communities of an ego and just analyze all of them, while one would expect that not all the communities have the same influence on the ego, but probably the most important are just one or two per ego. Or it could also be that the most influential users for the ego are not all concentrated in just one community

but are rather distributed in more than one.

For all these reasons, to get an answer on if and eventually how the most influential communities are related to the dynamics of the ego, we need more accurate analysis. If there is some effect, in our current analysis is certainly screened.

2.5 Paths

Leaving the refinement of the overlap analysis for further research, we now start to focus on the phenomenon of cascades which led to the collapse of iWiW. It has already been investigated [5] and it will the focus of the next chapter. In this section, we still concentrate on the rank correlation as an indicator, but instead of looking at the communities we now focus on some "cascading paths".

To construct a cascading path, we start from one of the 1600 egos we have already analyzed and look at all the neighbors who have a last login date prior to his. We select the neighbor that is "closest" to the ego in this sense, meaning the one who left just before him. Once selected the closest neighbor, we look at all his neighbors (not just the ones who are also linked to the ego) and apply the same selection process. We iterate this procedure for five step and we will end up with a chain of six nodes, ordered according to the last login date: the last one to have left is the ego and the first one is the last node of the chain.

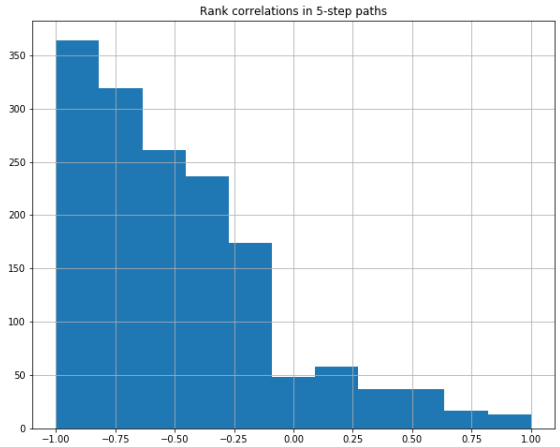


Figure 2.13: Rank correlation distribution in cascade paths

We do the same procedure for all 1600 egos and compute the rank correlation between the registration dates and the last login dates of the six nodes in the chains. We do the statistics of the collected coefficients, which is plotted in Figure 2.13. This time the result is much more evident than in the previous sections: there is a clear and very strong tendency towards anticorrelation, reinforcing the intuitive idea of the two opposite directions of the incoming and the outgoing dynamics that we mentioned before.

We see how by refining the analysis one ends up with clearer outcomes. In the previous

section, we were considering all the nodes in all the communities in the same way, even the largest communities where certainly any effect would be screened by the great amount of irrelevant information (it is hard to think that all the many nodes interact with all the others, which is actually more likely to happen in a small community). Although there was a slight tendency towards anticorrelation, the idea that we were trying to verify is much more evident in our last analysis.

In the next chapter, we will concentrate on the cascades, leaving these preliminary results about the dynamics and the similarity of communities and the relation between the flow of registering users and the cascade dynamics for future research.

Chapter 3

Cascades

We have seen some preliminary results about the dynamics of the social network and how the analysis so was not sophisticated enough to capture the role of aggregated metadata in the evolution of the service. We will now look deeper at the cascade phenomenon that brought iWiW to the collapse.

We want to identify the cascades and characterize them, looking at some of their properties. The first task is not as trivial as it seems, since in this case being an adopter for a node means leaving the service and in the end, all users left; therefore the network of adopters coincides with the whole network, a trivial and uninteresting result which does not say much about the fascinating endogenous effects and the cascading dynamics.

In this chapter, we develop a criterion according to which select those nodes who are more likely to have left due to the cascading effect, distinguishing them from the ones whose departure is most probably not related to the endogenous dynamics of the network. Once we apply the criterion to the entire network, we will have the set of "cascading nodes" and we will be able to reconstruct the structure of the cascades by going back in time towards the links between those particular nodes.

3.1 The criterion

We want to distinguish the users who most likely left the service due to the cascading effect from the ones that left randomly, due to exogenous effects, since it has been shown that the collapse has been caused by these two combined processes [5] (see Figures 1.10 and 1.11). In order to detect those users, we consider the number of friends who left in the last four weeks before the user's churning (m_{emp}).

Our criterion is based on the evaluation of the probability that, given the number of active friends W four weeks before the user's last login, m_{emp} of them leave in this four weeks period. The probability is computed according to the quadratically increasing counting process (linear rate) which represents the random departures (orange curve in Figure 1.10); given the hypothesis that users leave randomly with a time-linear rate, if this probability, summed to the probability of more extreme events (more than m_{emp} friends leave during the four weeks period), is very low, then we can consider that the node has

been involved in a cascading process. Formally, given the null hypothesis X that users leave randomly, we will evaluate the probability of the event Y that m_{emp} or more out of W friends have left in this time window (p-value). We will now formalize mathematically this concept.

Considering the random departures counting process X (null hypothesis) which grows quadratically with time (see Figure 1.10), we compute the probability of the event $Y =$ "m friends of the selected user left in the last four weeks before his departure (among the W friends still active at that date)" given the model X: $P_m(Y|X)$. Then we compute the p-value as the sum over m of all these events with $m \geqslant m_{emp}$ (m_{emp} is the empirical one). So basically, if the number of user's friends who left just before his departure is much greater than what is expected from the random departures model, then we can consider him a "cascading leaver".

To compute this value we use the following parameters:

- Constants:
  - τ (last login date of the selected user)
  - T (4 weeks)
  - m_{emp} (empirical number of friends who left in the interval $[\tau - T, \tau]$)
  - W (total number of active friends of the selected user at $\tau - T$)
  - M (total number of active users in the social network at $\tau - T$)
- Variables:
  - $\mathbb{E}[N(t)] \equiv N(t)$ (total expected number of inactive users at time t (hence the number of users that have left before time t) according to the random departures model: the quadratic orange curve in Figure 1.10)
  - $n(t)dt = \frac{dN}{dt}dt$ (number of users who leave in the interval $[t, t + dt]$)

Given τ (and hence M), $N(t)$ grows quadratically with time and so does $N(t)/M$, which is the probability that a generic node leaves before time t (considering that the process starts at time $\tau - T$); $n(t)/M$ is the probability density associated with this process. The probability we are looking for is the probability that m nodes (among W) leave in the interval $[\tau - T, \tau]$ and the other $W - m$ leave after time τ, among all the possible combinations:

$$P_m(Y|X) = \left[\frac{\int_{\tau-T}^{\tau} n(t)dt}{M}\right]^m \left[1 - \frac{\int_{\tau-T}^{\tau} n(t)dt}{M}\right]^{W-m} \frac{W!}{m!(W-m)!}$$

And integrating:

$$P_m(Y|X) = \left[\frac{N(\tau) - N(\tau - T)}{M}\right]^m \left[1 - \frac{N(\tau) - N(\tau - T)}{M}\right]^{W-m} \frac{W!}{m!(W-m)!}$$

which is actually a binomial probability distribution. If we define $P_m(Y|X) \equiv P_m$ and $q \equiv \frac{N(\tau)-N(\tau-T)}{M}$ we obtain:

$$\begin{aligned} P_m &= q^m[1-q]^{W-m}\frac{W!}{m!(W-m)!} \\ &= \binom{W}{m} q^m[1-q]^{W-m} \end{aligned} \tag{3.1}$$

Of course, q (and hence P_m) depends on τ (the last login date we are considering). We can see how the shape of the distribution P_m changes with time in Figure 3.1, where we plot an example with $W = 200$ friends still active at $\tau - T$, for different τ.

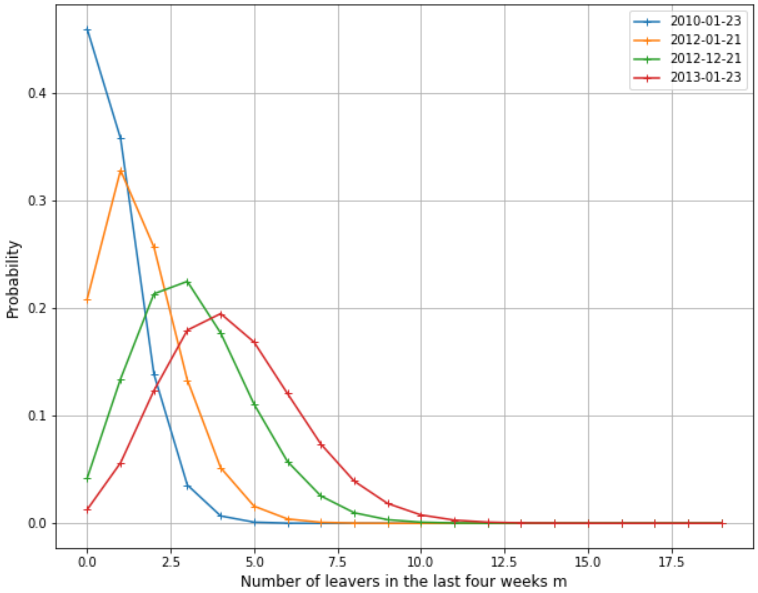


Figure 3.1: P_m shape for different τ

We notice how the shape gets broader with time and how larger m values get a finite probability as time passes: this makes sense, as the hypothesis X we are testing is based on a counting process whose rate grows linearly with time, so the more time passes the more people are expected to leave the service (which is also true for a time window like T). Hence for later times, it becomes more likely that many friends leave during the last four weeks before the user's departure, while the probability of the same event is cut off for earlier times.

To obtain a criterion according to which select the "cascading nodes", we consider both the probability of the observed event and the probability of more extreme events ($m > m_{emp}$). Referring to Figure 3.1, we consider the right tail of the distribution, summing over m and obtaining the p-value p:

$$p = \sum_{m \geqslant m_{emp}} P_m$$

If p is very small, it means that our hypothesis X of random departures does not describe well the observed event and is therefore rejected, which is equivalent to say that if the observed number of friends' churnings is much greater than the expected one, we can consider the selected user's departure a cascading event. We fix the threshold for p at 0.05, meaning that if for a user we find that $p < 0.05$, we consider him a cascading node.

In Figure 3.2 we show how the p-value changes with time for different m_{emp} (with $W = 200$ like in Figure 3.1). The meaning of the figure is clear: when the p-value curve is under the threshold line $y = 0.05$, we consider the respective event a cascading event. We see how for small m_{emp}, the p-value crosses the threshold line $y = 0.05$ very early (or does not), meaning that few friends leaving just before the user's churning can be considered a cascading event only for early dates (where fewer people were leaving the service, according to the linear rate). On the other hand, the departure of more friends in the T period is taken as a cascading event also at larger times.

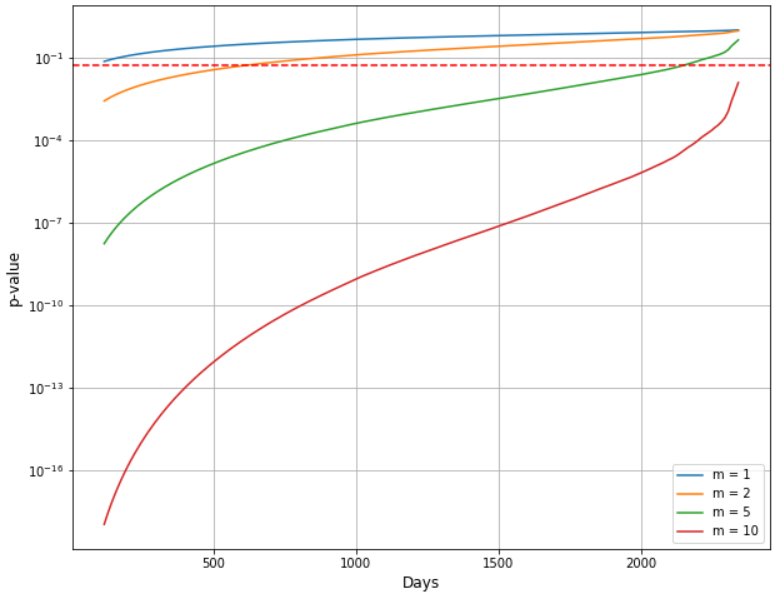


Figure 3.2: Evolution of the p-value (in semi-logarithmic scale) $W = 200$ during the social network's life. The different curves represent different m_{emp}. The horizontal dashed line is the threshold $y = 0.05$.

Now we developed a method to select the cascading nodes, which we will use to reconstruct the entire cascades back in time and then characterize them.

3.2 Aggregate results

We apply the method to all the 4298776 users with available last login date and timestamped acquaintances. The result is shown in Figure 3.3.

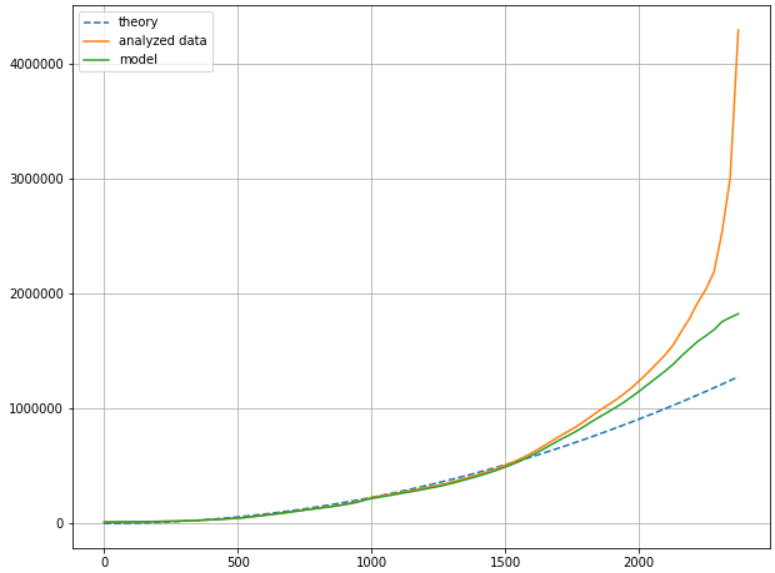


Figure 3.3: In orange, the evolution of the total number of inactive users (the same as the blue curve in Figure 1.10. In blue (dashed), the parabola (orange curve in Figure 1.10). In green, the users with p-value > 0.5 (non cascading).

The orange and the dashed blue curves are the same of Figure 1.10. We added the result of our "filter": the green line is the evolution of the total number of users whose p-value was found to be greater than 0.05 and, therefore, who we will not consider as cascading nodes. The difference between the orange and the green curve is the evolution of the total number of cascading nodes.

The green curve does not overlap very well with the dashed curve, but this is not surprising: it is true that we took as the hypothesis to reject the dashed curve (linear rate), however the choice of the threshold 0.05 for the p-value is completely arbitrary; eventually, if one wished to have a better overlap, he would just need to increase the threshold. Also, the dashed curve is just a model for the random churners, based on the initial shape of the real inactive users curve; but this does not guarantee the future evolution of the number of random churners to be the same as before the bifurcating point (where the two curves of Figure 1.10 separate). And finally, even the choice for $T = 4$ weeks is arbitrary (another choice would change the green curve).

However, the qualitative behavior of our filtering is what we were expecting and looking for: initially, there are almost no cascades; then, after the bifurcating point, the two curve

start to separate, meaning that more and more cascades are triggered as time passes.

Of the total number of users analyzed, 57.5% have been selected by our criterion as cascading nodes.

3.3 Identifying the cascades

Now that we have identified all the cascading nodes, we can build the cascades, which are subgraphs of the original network which we construct in the following way.

We start from a cascading node and we look at the neighbors which are cascading and which left in the four weeks period before our node's departure (we will refer to the neighbors which have those two properties as "cascading neighbors"). Once selected, we add edges (in the cascade graph) between them and the starting node. Then, we do the same for the selected neighbors, looking for their cascading neighbors, but selecting only nodes that have not been added to the cascade graph yet, in order to avoid loops and end up with a tree structure.

A fictitious but useful example is shown below (both the network and the selection of the cascading nodes are random): in Figure 3.4 we show a network where the node from which we start building the cascade (the root) is in red. The nodes in light blue are the ones that have been selected to be part of the red node cascade, either because they are cascading neighbors of the red node or because they are cascading neighbors of the cascading neighbors of the red node etc... The nodes in blue are all the other nodes (note that they might also be cascading nodes, but may have left after the node to which they would be linked or more than four weeks before). Finally, we end up with the tree structure in Figure 3.5, which is, according to our criterion, the cascade structure that brought to the churning of our selected red node.

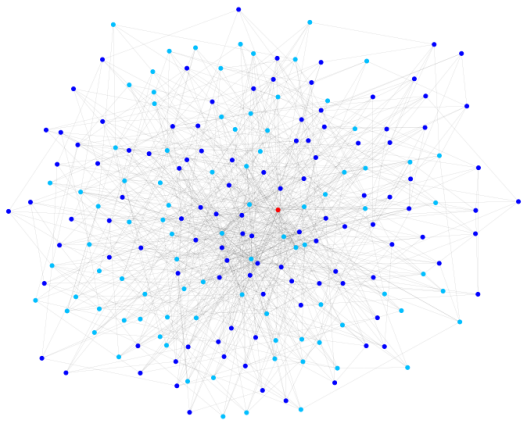


Figure 3.4: Fictitious network: starting node in red; in light blue, nodes that have taken part in the (fictitious) cascade of the starting node; in blue, all other nodes.

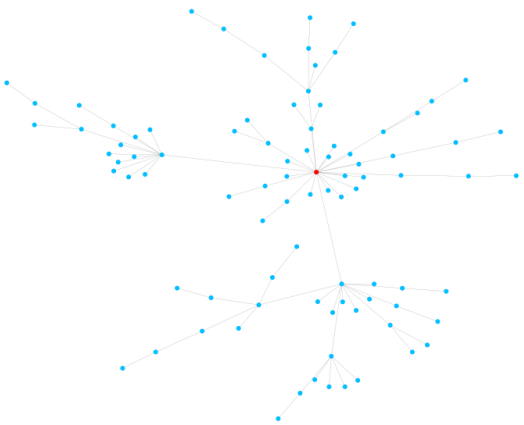


Figure 3.5: The fictitious cascade tree of the starting node (red), obtained from the fictitious graph in Figure 3.4.

3.4 Results

We are now ready to apply our methods to the data: starting from a cascading node, we can build his cascade tree, which, according to our model, is the endogenous dynamics that caused his departure. We remind that in a cascade tree, two nodes are linked to each other if they are cascading nodes, if their last login dates are closer than four weeks and, of course, if they are linked in the iWiW network.

Since the number of selected cascading nodes is very high (57.5% of the total number of users, which means almost 2.5 million nodes), identifying the cascades takes quite a long computational time. We present here the results for ten nodes, chosen at random among the cascading nodes.

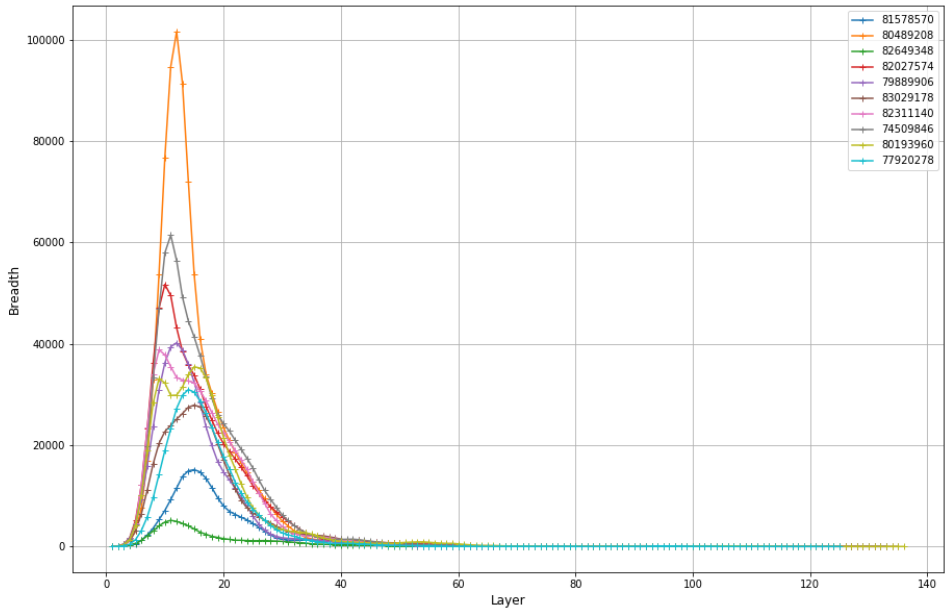


Figure 3.6: The breadth of each layer of ten different cascades. The label in the legend is referred to the I.D. of the starting node (the root of the cascade).

The shape of the cascades is shown in Figure 3.6: every curve is one of the ten analyzed cascades. We said that the cascade graphs we built are trees and hence, taking the starting node as the root, every node of the structure belongs to a well-defined layer (or generation). In Figure 3.6 we are plotting the number of nodes belonging to each layer (the breadth of each layer).

There are some interesting characteristics which stand out. First of all, we notice how the largest layers are between layer 5 and layer 30, meaning that the majority of the cascade nodes is concentrated in this region, independently of the starting node. Secondly, while

the very first layers and the last ones seem to have approximately the same size for the different cascades, the interesting region where most nodes are placed shows a very rich behavior: in fact, the maximum breadth ranges from ∼5000 to ∼100000 nodes, outlining very large fluctuations.

It is interesting that despite reaching the maximum breadth in the initial layers, the cascade can reach a very long tail, meaning that the initial trigger of the cascade (in the last layers) can be very far in time from the last login of the starting node. Indeed, the dynamical process might be quite long in time. We plot a detail of the shape of the last layers in Figure 3.7, showing that also in this layers' region one can find a quite rich behavior by looking at a different scale.

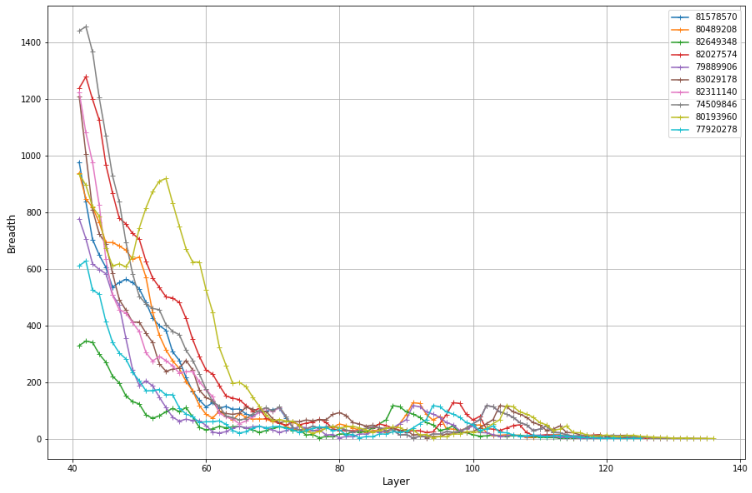


Figure 3.7: Zoom on the tail of 3.6.

As expected, the cascades can be very large and hence computationally hard to detect. We expect that the size of the cascades (meaning the total number of nodes) depends on the last login date of the starting node we are looking at. In fact, if the selected node left the service late, there are more nodes who might potentially be part of his cascade. Indeed, this is the case, as shown in Figure 3.8: the cascade size grows monotonically with the last login date of the selected user.

Since we selected the ten starting nodes at random and since the size of their respective cascade might be very large, we are interested in seeing whether the cascades overlap, i.e. if one among the ten selected starting nodes also belongs to the cascade of some of the other nine. What we find is find is at the same time expected and surprising: the cascades do not simply overlap in one or a couple of cases, but they are actually all parts of one single cascade. We would have expected, given the large fraction of cascading nodes selected by our criterion, that by analyzing all the cascades, we would have found that the great majority of them, if not all, are all parts of one single giant connected component. Instead, we find that this is the case just by picking at random ten cascading nodes.

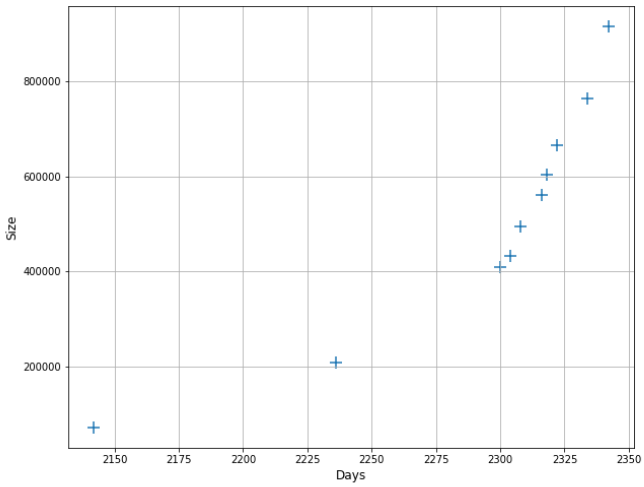


Figure 3.8: Size of each of the ten analyzed cascades as a function of the last login date of the cascade root.

To give an idea of the overlap we plot the results in the directed network in Figure 3.9. The nodes are the ten roots selected at random whose cascade shapes are plotted in Figure 3.6. An arrow from node i to node j means that i belongs to the cascade of j. Of course, a link has a single direction as one of the two nodes left the OSN later than the other and hence it can not belong to his cascade. Also, one can check that if i belongs to the cascade of j, who belongs to the cascade of k, then also i belongs to the cascade of k.

3.5 Discussion

We have developed a criterion which allows us to select the users who have left the service as a result of a cascading effect and to distinguish them from the random churners. There are some arbitrary parameters in the criterion, like the four weeks timescale or the p-value threshold at 0.05, that are crucial for the final configuration of cascading nodes. Our choice gives us the qualitative outcome we were looking for: the cascades start to emerge at a certain time (bifurcation point) and their number starts to grow after that time.

With this criterion, we had the possibility to reconstruct the shape of a few cascades by looking at the entire network topology, the last login dates of each node and the configuration of the cascading nodes. The resulting cascades are tree graphs where the leaves are the first churners (the triggers of the cascades) and the root is the last churner (the selected node from which we started building the whole structure). This graph gives us an idea of the dynamical structure which led to the churning of the root node.

The structures of the cascades show some interesting features: while there is a common rapid growth (in terms of number of nodes) in the first layers and very long and narrow branches, the region between layer 5 and layer 30 shows some very rich behavior and large

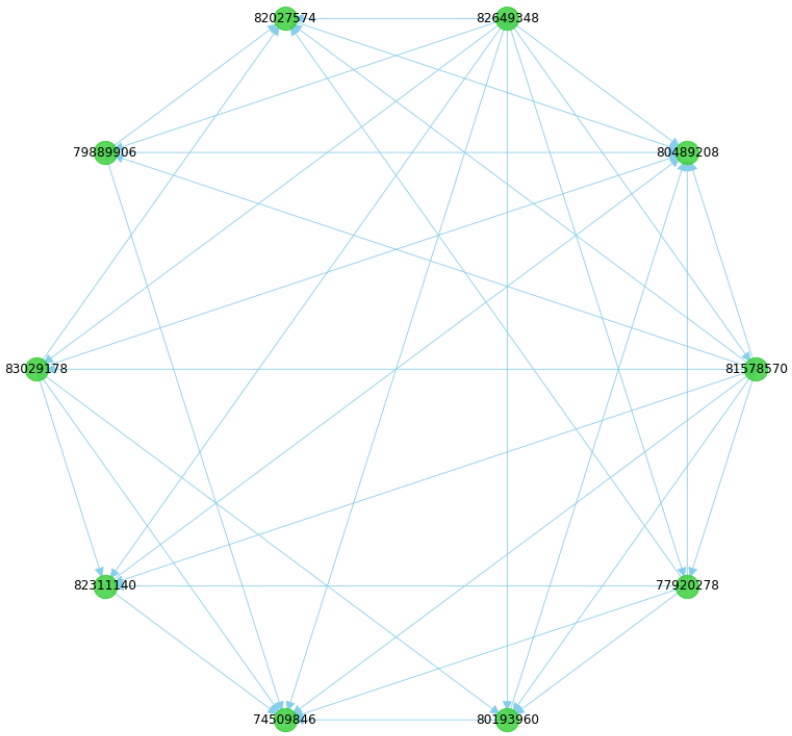


Figure 3.9: Graphical representation of the overlap between the analyzed cascades. The node at the tail of an arrow belongs to the cascade of the node at the head of the arrow.

fluctuations.

The size of the cascades increases with the last login date of the root node, as expected from the observed growth of cascading nodes with time. Finally, we observed that the analyzed cascades have a strong overlap, being each a portion of the same single cascade.

Chapter 4

The model

We have already seen how threshold effects in social networks like iWiW trigger global cascades which may generate the failure of such systems. In this chapter, we build a very simple dynamical model in order to understand how the parameters related to growth and to the shrinking of the network bring the system size to equilibrium. We generalize the Watts and the Kertesz models [6, 14] in order to apply it to the evolution of the structure of online social networks (OSN) and see when and how those systems collapse due to cascading phenomena and if these effects can be balanced by a constant growing rate of people joining the platform.

4.1 The iWiW case and the model

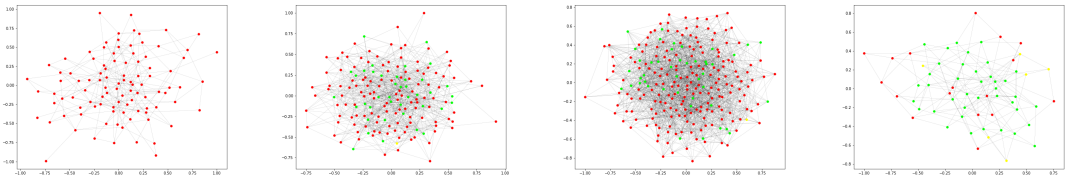
We have already seen how the combined dynamics of random and cascading churners is what seems to have generated the collapse of iWiW. We remind with the help of Figure 1.10 how the two different factors mainly drove the failure dynamics of the OSN: initially, departure from the network is mainly due to a linear rate of random churners, while at some point cascades start to dominate the dynamics and they rapidly bring the network to the collapse.

Since we assume that these issues are valid for any OSN, we focus here on the reasons why some of them manage to survive. Our idea is to discuss whether introducing a growing or recovery rate in the dynamics of the system is enough to make it survive. We assume that apart from people leaving the network for exogenous reasons (random churners) and people churning for endogenous issues (threshold) generating the departure cascade, at any time new people decide to join the service for the first time or people who have left decide to join again.

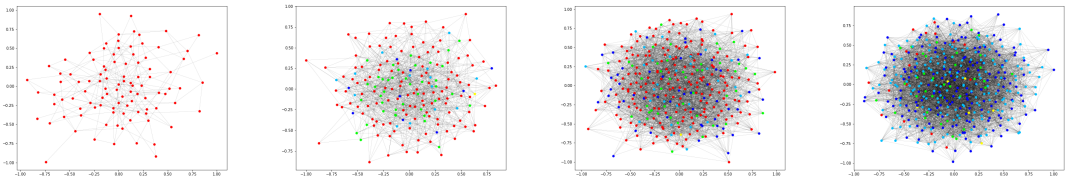
To model such a system, we consider three parameters which drive the dynamics of the OSN: the probability p that at any time a node leaves spontaneously, which is responsible for random departure (in the iWiW case, the random churning rate is growing linearly with time, but for simplicity we consider a constant rate p); the individual threshold ϕ, which is the main responsible for cascading failures; the growing rate k.

We consider a sample of N individuals, N_{in} of which have already joined the social

network. At each time step, inactive individuals join the network with rate k and active users leave the network with rate p or because a fraction $x > \phi$ of their neighbors have already left. We let the system evolve until it converges and take $\rho =< N_{active} > /N$ as order parameter, where N_{active} is the number of active users at a given time step and the average value is taken over the last 100 steps of the simulation if the system hasn't collapsed, otherwise it is simply zero. To clarify the evolution of the system, in Figure 4.1 we plotted the steps of the evolution with a specific set of parameters.



(a) Network of active users



(b) Network of all users who have been in the system, both active and inactive

Figure 4.1: Evolution of a simulation with $p = 0.1$, $\phi = 0.4$ and $k = 0.05$: from left to right, time steps 1, 3, 5 and 7 (the last before the collapse). Red nodes have joined at previous time steps; green nodes are new incomers; blue nodes have left the service at the previous time steps; light blue nodes have just left the service; yellow nodes (very few in this case) are inactive users who join the service again.

In the figure, the above list of images is the time history of the service (active users), while the list below is the evolution of the network of both active and inactive users, which clarifies the way nodes leave the service as time goes on. We notice how initially all nodes are active and how in the first steps of evolution some users join the system while the parameter p is responsible for some initial random departures. Going forward with the evolution, the number of inactive users starts to be consistent enough to make threshold departures rise. At this point, we can appreciate how these effects have a dramatic effect on the service number of active users which decreases very rapidly and how the new incomers rate is not enough to avoid the system to collapse. We also notice that some (very few) of the previously inactive users join the system again (yellow nodes). Despite the simplicity of the model, the dynamics of the extremely fast collapse looks qualitatively similar to what has happened to iWiW.

We want to understand if and how the model we constructed could prevent collapse by tuning the parameters which govern the dynamics.

4.2 Results

We make a set of ten simulations of 1000 time steps for each pair of values k and ϕ (we take the parameter p fixed to 0.1) and analyze the behavior of $\rho =< N_{active} > /N$. The simulations show that the system undergoes a discontinuous phase transition (Figure 4.2) with a very clear interpretation: in order for the system to prevent cascading collapse, it has to keep a minimum finite size. If this is not the case, the growing rate k is not enough to provide the system a stable finite size and the threshold effect rapidly overcomes. We will refer to the two different phases as the collapsing phase and the stable phase.

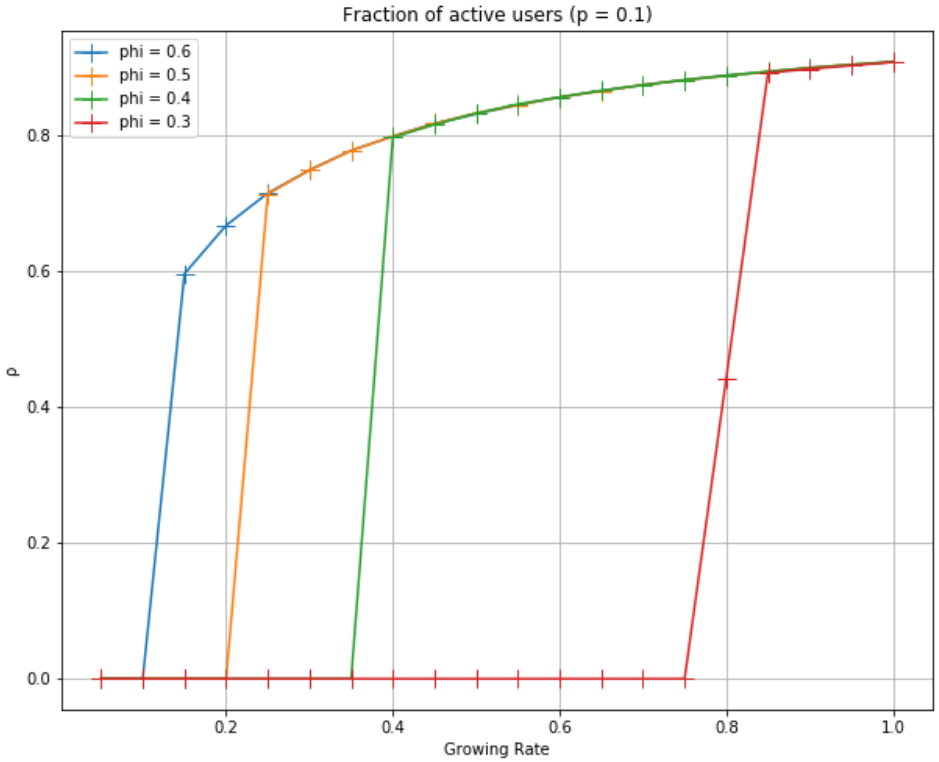


Figure 4.2: ρ dependence on k for different ϕ values.

The critical value k_c grows by decreasing the threshold parameter ϕ, as expected: a low threshold makes users leave just with a few inactive acquaintances, hence accelerating the cascading collapse; in this case, a higher growth rate is necessary to keep the system alive. Also, the minimum finite (normalized) size $\rho_{min}(\phi)$ is reached at the critical point k_c, meaning that beyond that point a higher growth rate keeps the system stable with a larger fraction of users. These are all intuitively expected results.

It is interesting to observe that the different curves overlap perfectly when they reach

the stable phase, no longer depending on ϕ: this means that once the system has managed to balance the threshold effects, his size depends only on the growing rate k and we may assume that cascades have no influence in the stable phase. To verify the latter statement, we made simulations over the whole range of values in the (ϕ, k) plane and plotted the results in the heat map in Figure 4.3.

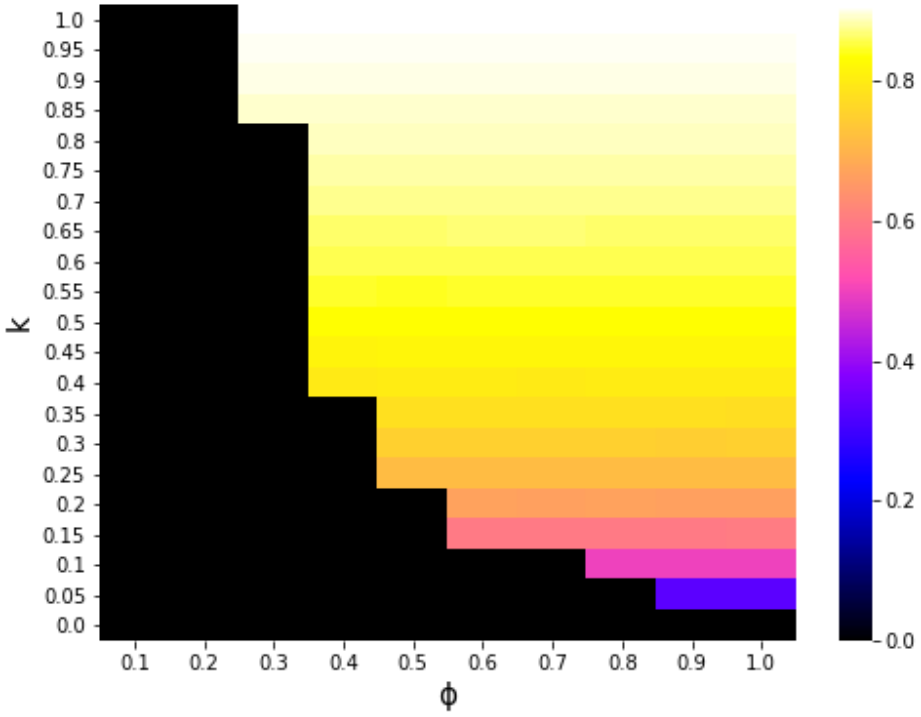


Figure 4.3: ρ heat map in the (ϕ, k) plane.

All the previous analysis are confirmed. The system clearly undergoes a first-order phase transition with a critical value k_c decreasing with ϕ. In the stable phase, the color changes moving vertically but not horizontally, confirming that the phase does not depend on ϕ. It is worth observing how even for low ϕ values the system has a stable phase, meaning that even if cascading effects are triggered very easily they can be balanced by a sufficiently high growth rate. Nevertheless, there is a region $\phi < 0.3$ where no k value is able to balance the cascades and the service necessarily collapses. This portion of the plane would certainly vary with different p and we suppose that by decreasing it the stable region would be larger and maybe reach the $\phi < 0.3$ region.

As suggested by the plot in Figure 4.4, k_c seems to follow a power law as a function of ϕ with critical exponent $\alpha = 2.38$.

We also expect the time required for the service to collapse to follow a diverging scaling

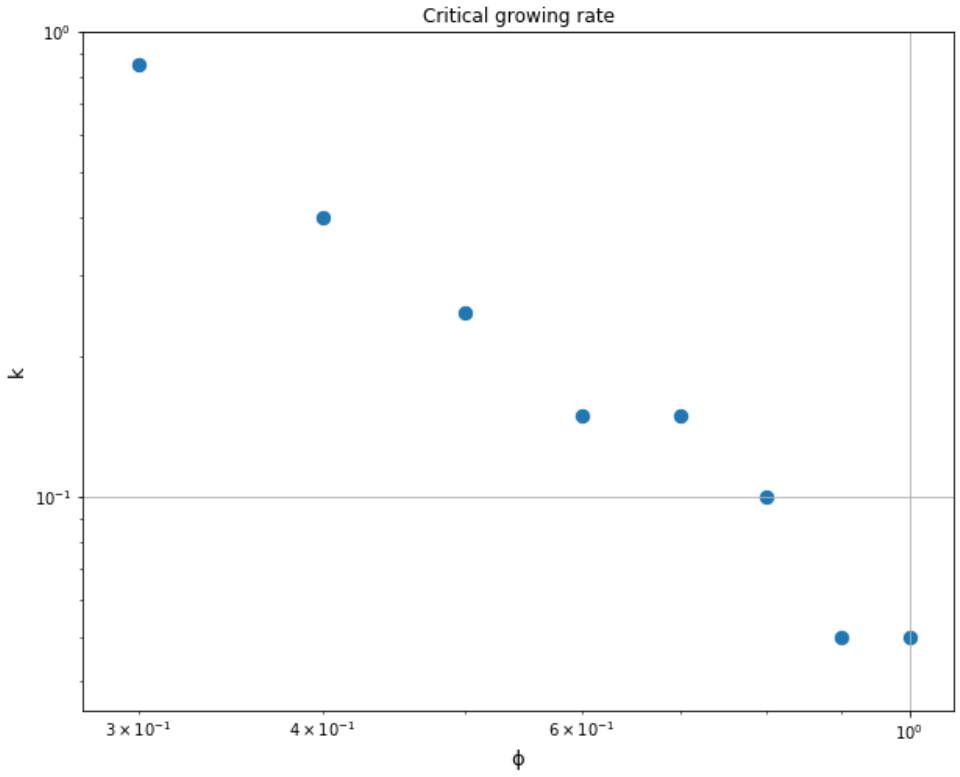


Figure 4.4: Scaling behavior of the critical value k_c as a function of ϕ.

law behavior when the parameters approach the critical point, as fluctuations usually show this kind of diverging behavior in critical phenomena. The latter assumption is motivated by observing that when approaching k_c the system manages to survive for a longer time until a large fluctuation suddenly breaks the apparent equilibrium (metastable state) and rapidly drops ρ to zero (see Figure 4.5).

This scaling divergence seems to be confirmed in Figure 4.6, where we plotted the collapsing time as a function of k near the critical point for a particular choice of the parameters, but the result is also valid for other values (not shown).

4.3 Non-interacting mean field model

The simulations show very clearly a discontinuous phase transition. We want to make sure that this result is confirmed at least qualitatively by an approximated analytical model of the system. We first notice in Figure 4.5 the temporal evolution of three simulations near the critical point: we can see how initially the dynamics is mainly driven by k and p since at early time steps there are too few inactive users to trigger global cascades; after some

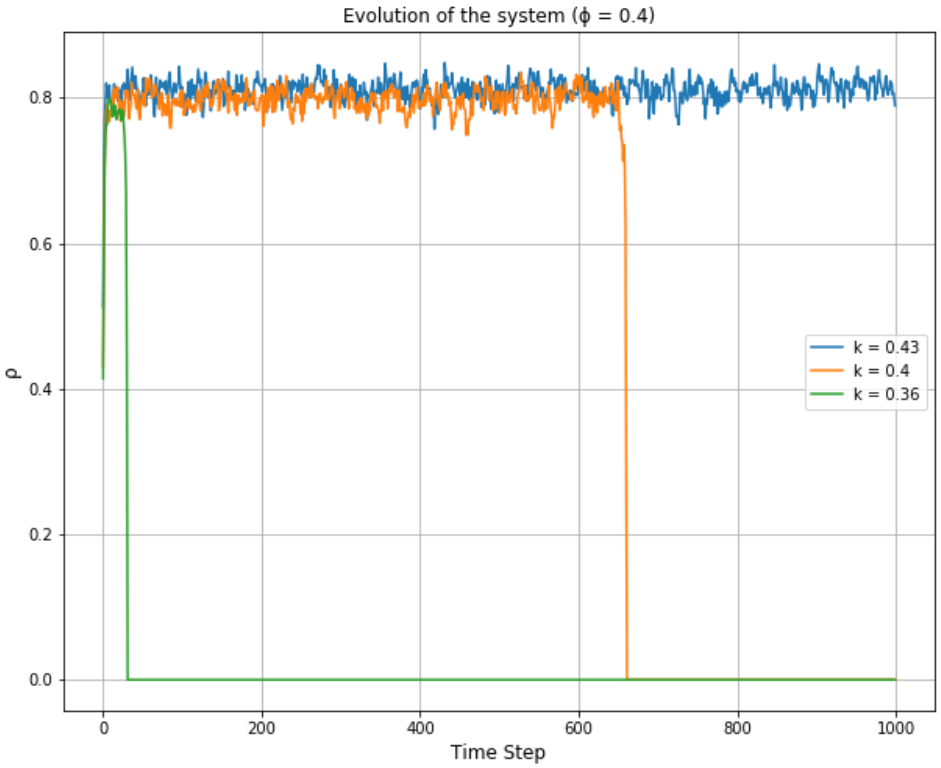


Figure 4.5: Evolution of ρ in time during some simulations near k_c for $\phi = 0.4$.

time, a consistent fraction of the N potential users have joined and left the system and at this point cascades may be triggered or balanced. Ideally, the evolution is divided in two distinct dynamical phases: initially the system reaches a metastable state due to the effects of only k and p (non-interacting dynamics); then, the threshold ϕ determines whether the system may collapse or survive, based on the global properties of the metastable state it has reached (mean field dynamics).

With these assumptions in mind, we first look at the expected number of active users step by step in the non-interacting dynamics. Starting with n_0 users, after one time step the expected number will be $n_1 = n_0 + k(N - n_0) - pn_0$. Since at each time step the dynamics is the same, rearranging the terms we can write the recursive relation:

$$n_i = kN + (1 - k - p)n_{i-1}$$

Solving the recursive relation and defining $x = 1 - k - p$ we obtain the relation for the expected number of active users after t time steps, starting from n_0:

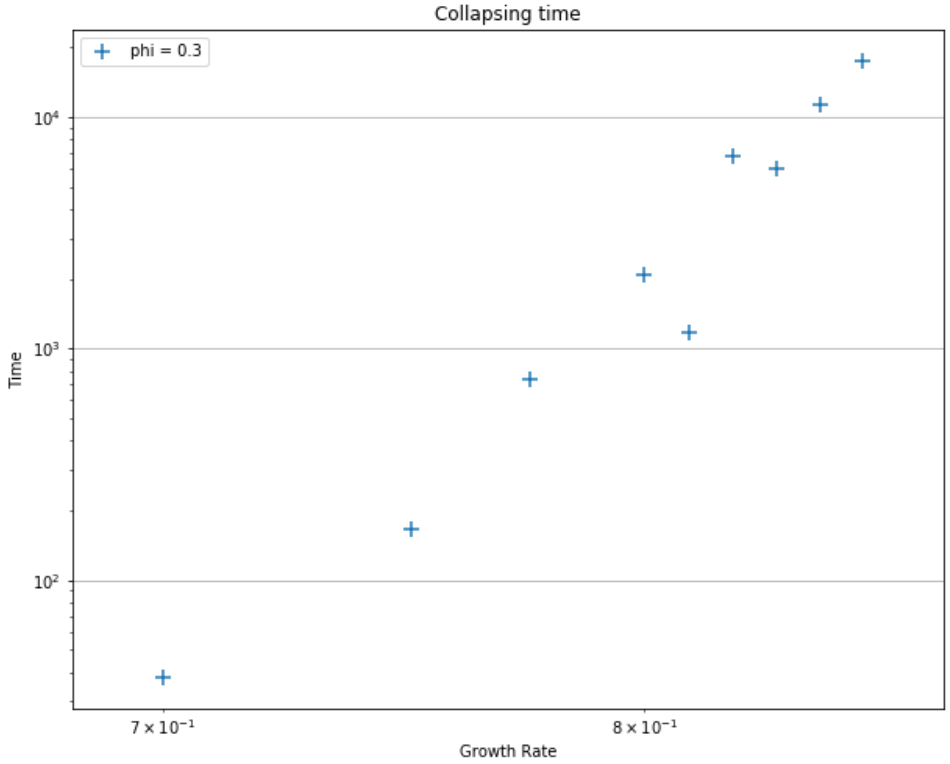


Figure 4.6: Scaling behavior of the collapsing time.

$$n_t = kN \sum_{i=0}^{t-1} x^i + n_0 x^t$$

The metastable state will be obtained in the limit of infinite t. Reminding that $\mid x \mid < 1$ we obtain:

$$n_\infty = kN \frac{1}{1-x}$$

which corresponds to a density:

$$\rho_\infty = \frac{k}{1-x} = \frac{k}{k+p}$$

Now the second part of our simplified dynamics comes over. We make a very simple mean field approximation and consider the final network as fully connected: in this way, the global density ρ_∞ will coincide with the local fraction of active nodes of every user, which will be the same for everyone. Hence the local threshold ϕ which should be applied

to every single node now becomes a global property of the system: if the fraction of inactive users $1 - \rho_\infty$ is greater than ϕ the system collapses (since the threshold condition is satisfied for every node), otherwise it survives with the finite size n_∞. With these simple approximations, the collapse condition becomes:

$$\phi < 1 - \frac{k}{k+p}$$

In Figure 4.7 we plot this condition and the critical points of the simulations (the ones of Figure 4.3). The mean field model exhibits a discontinuous phase transition as well and the shape of the critical curve between the two phases is the same. Nevertheless, the empirical points don't fit the curve because of the brute mean field approximations. The distance between the points and the mean field curve becomes greater with small ϕ and this is also expected. Indeed, with our approximations the system evolves regardless of the threshold condition until it reaches ρ_∞; in reality, the parameter ϕ enters in the dynamics mechanisms the earlier the smaller it is.

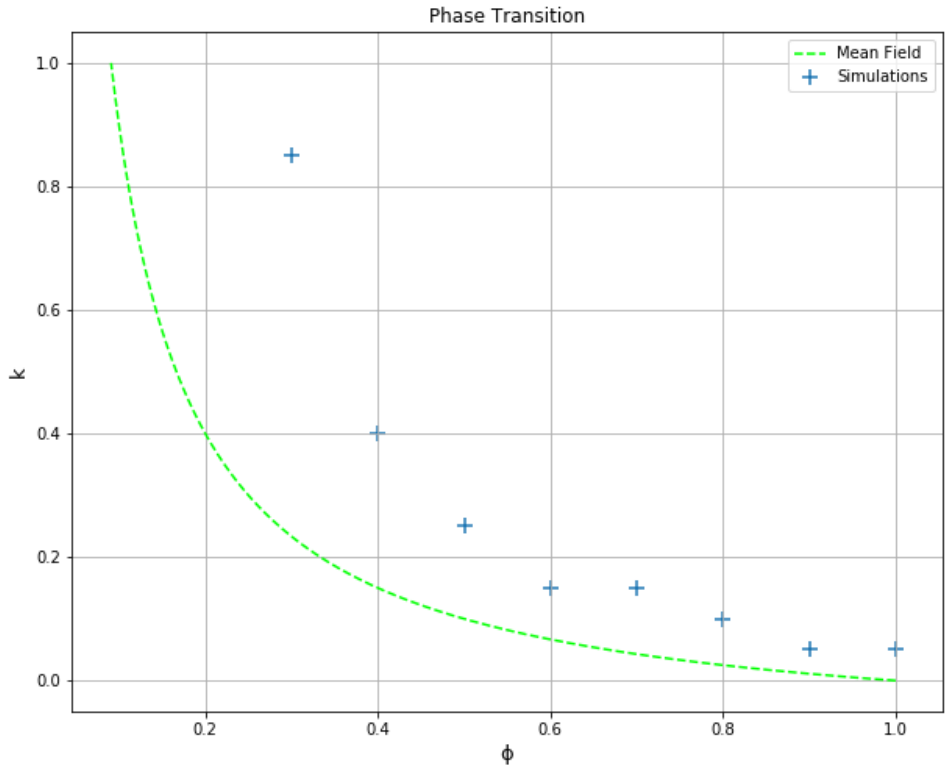


Figure 4.7: In green, the phase transition predicted by the non-interacting mean field model. The scattered blue points are the phase transition obtained from the simulations.

Our analytical model approximations are too strong to make some predictions about the system as it is only able to show the shape of the transition. Much more refined approximations are required to have a better fit with the simulations.

4.4 Discussion

Every online social network' s survival is based on how many people register and how many leave. Some platforms such as Facebook and Twitter survive and grow in term of active users while some others collapse, as it happened to iWiW. It is very important for such services to understand which parameters will determine their future dynamics while they are still active. However, each network's dynamics is governed in a non-trivial way by many and different endogenous and exogenous factors such that each OSN should be considered and analyzed separately.

Aware of these limitations, we made a general simple model and from the simulations we found that by tuning the growth rate k of new registered users and the threshold ϕ which is responsible for endogenous cascading collapse the system undergoes a discontinuous transition between two distinct phases in the (ϕ, k) plane: a collapsing phase, where the system collapses, and a stable phase, where the system is able to keep a finite size. Both the critical value k_c as a function of ϕ and the collapsing time near the critical point follow power laws, reinforcing the phase transition result.

We tried to model the behavior of the system by splitting the dynamics into two distinct parts: a non-interacting evolution which brings the system to a metastable state and a mean field one, governed by the threshold ϕ, which drives it to the equilibrium state (collapse or finite size). Although the bad fitting with the results of the simulations due to the strong approximations, we find the same qualitative behavior (discontinuous phase transition in the (ϕ, k) plane).

In the iWiW case, it has been shown how a very simple generalized threshold model is able to fit the data of the collapse very well[5]. Even though the dynamic was different (the random departure rate p was linear with time instead of constant, and there was another parameter relative to waiting time of a user between threshold overtaking and effective departure), our results allow us to make at least some qualitative conclusions.

According to our model, the system could have survived with an appropriate growing rate k; in fact, even with a low threshold value, the stable phase of our model occupies a non-vanishing portion of the plane (even larger with a smaller p). These assumptions seem to be confirmed by Figure 4.8: considering the time evolution of the registration rate (the analogous of our k), we can observe how besides an initial growth it decreases with time until it reaches zero. This means that at a certain point the growth rate reached a critical value without increasing again and the network was not able to survive any longer. At the same time, we see the limitations of our model, where we considered a constant k while in the iWiW case its evolution is much more complex and should be considered in future research. What drives the dynamics of the registration date could also be very interesting, but it's a completely different research topic out of our current purposes.

For the sake of simplicity, we considered a constant p during the simulations, while in the iWiW case it showed a linear dependence on time. It would very interesting to make more sophisticated simulations in order to have a better model of the empirical observation

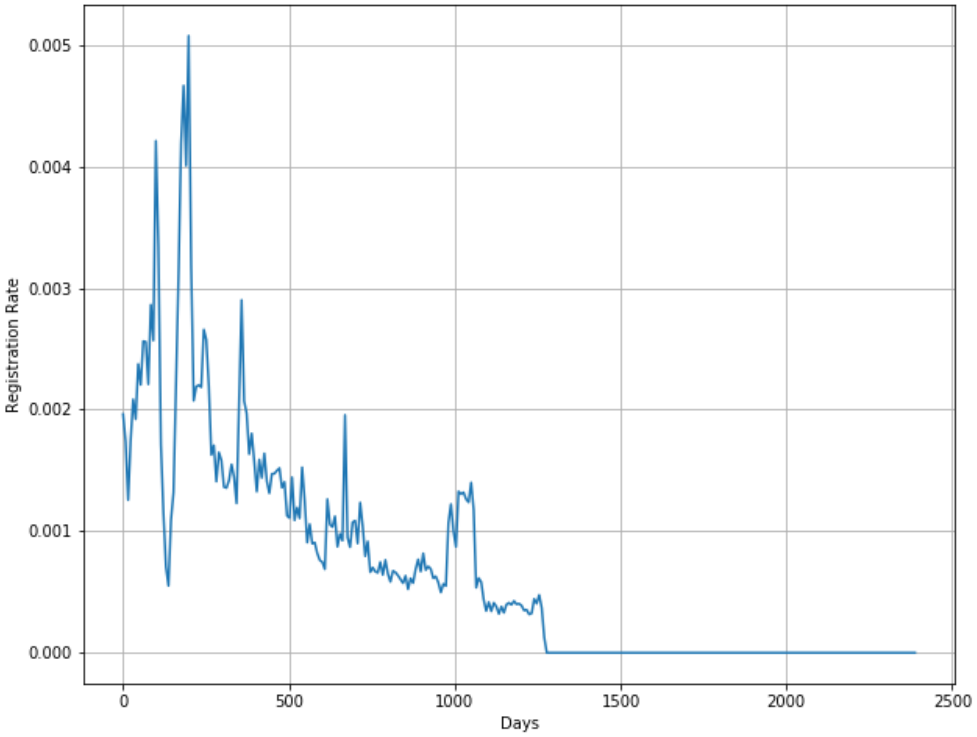


Figure 4.8: Evolution of the registration rate in time (iWiW data)

and, if available, to compare them with other OSNs data.

Despite its simplicity, we think that our dynamic model is able to explain various real-world scenarios of both OSNs or non-virtual aggregation platforms such as affiliation clubs or communities. Although the model is not able to incorporate all the nonetheless important endogenous and exogenous factors in an actor's decisions, it captures some of the main dynamical features. We hope that future analysis will help in the direction of modeling and predicting such dramatic cascading failures.

Chapter 5

Conclusions

Cascading phenomena in social networks and in other real systems can lead to unexpected and dramatic outcomes, like the financial contagion following a shock in a network of banks, the rising of a riot or the cascading failures in infrastructures. In this research we have tried to get a better understanding of the dynamics of such effects in a social context by analyzing the data of the Hungarian online social network *iWiW*.

Our aim was to have an overview of how a user becomes an adopter in the cascade which has been triggered in the social network (where in this case, *adopting* stays for *leaving the service*). We first analyzed the dynamics inside the communities which formed around a node (an ego) and we found out that the registration and last login dynamics of the users inside a community are not random. We then introduced a measure of similarity (the overlap) between a community and the respective ego to see whether these to quantities were related, in order see whether there is a common pattern in the most overlapping communities (which are thought to be the most influential in the ego's decisions). We need more accurate research to address this issues, as our level of analysis is not able to provide a satisfactory answer.

Instead of focusing on the communities which formed around the egos, we looked at the dynamical features of some "cascading paths" and measuring the rank correlation between the registration and the last login dates of the chained nodes we found a very strong tendency towards anticorrelation (the first to leave is the last who has registered).

Then we managed to develop a criterion according to which select the users who left the service due to endogenous effects (the ones who triggered the cascade failures) distinguishing them from the random churners who left due to exogenous effects, as it has been shown that the failure dynamics of iWiW was mainly driven by these two types of dynamics. Selecting the cascading users according to this criterion, we were able to reconstruct the entire cascade structure which led to the abandon of a given user and gives us an idea of the cascading dynamics.

Finally, inspired by the dynamics of iWiW we developed a generalized Watts model to see whether threshold effects can be balanced by a proper growing rate of new incoming users. We found out that the system undergoes a first order phase transition in the parameter space with two distinct regimes: a collapsing phase, where the threshold dynamics overcomes and leads the system to the collapse, and a stable phase, where the rate of new incomers balances the cascading effects and the system reaches an equilibrium finite size.

We have investigated the phenomenon of cascades in the breakdown of iWiW, adding some findings to the previous knowledge about the service and pointing out new challenging questions for the future research. In particular, the next goal is to investigate the role of communities in the cascades we identified, in order to obtain coarse-grained cascades, and to use the metadata to characterize the qualitative structure of the cascades.

Bibliography

[1] M. Newman, *Networks: an introduction*, Oxford University Press, 2010.

[2] M. Newman, *The structure and functions of complex networks*, SIAM review, 2003.

[3] J. Travers, S. Milgram, *The small world problem*, Psychology Today, 1967.

[4] J.L. Moreno, H.H. Jennings, *Who shall survive?*, Nervous and Mental Disease Publishing Co., 1934.

[5] J. Kertesz, J. Török, *Cascading collapse of online social networks*, Nature, 2017

[6] D.J. Watts *A simple model of global cascades on random networks*, Proceedings of the National Academy of the United States of America, 2002

[7] M. Granovetter, *Threshold models of collective behavior*, American Journal of Sociology, 1978

[8] T.C. Schelling, *Dynamic models of segregation*, Journal of Mathematical Sociology, 1971

[9] P.A. Dow, L.A. Adamic, A. Friggeri *The anatomy of large Facebook cascades*, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013

[10] J. Cheng, J. Kleinberg, J. Leskovec, D. Liben-Nowell, B. State, K. Subbian, L. Adamic *Do diffusion protocols govern cascade growth?*, Association for the Advancement of Artificial Intelligence, 2018

[11] D.J. Watts, E. Bakhsy, J.M. Hofman, W.A. Mason *Everyone's an influencer: quantifying influence on Twitter*, Proceedings of the Fourth International Conference on Web Search and Data Mining, 2011

[12] P.A. Dow, L.A. Adamic, J. Cheng, J. Kleinberg, J. Leskovec *Can cascades be predicted?*, Proceedings of the Twentythird International Conference on World Wide Web, 2014

[13] B. Lengyel, R. Di Clemente, M. Gonzalez, J. Kertész *Spacial diffusion and churn of social media*, arXiv:1804.01349, 2018

[14] Z. Ruan, G. Iñiguez, M. Karsai, J. Kertész *Kinetics of social contagion*, Physical Review Letters, 2015

[15] B. Lengyel, A. Jakobi *The offline landscape of an online social network: distance and size shaping community spread and activity*, 52nd Congress of the European Regional Science Association, 2012

[16] M. Bluhm, E. Faia, J.P. Krahnen *Endogenous banks' netowrks, cascades and systemic risks*, Mimeo Goethe University Frankfurt, 2012

[17] M. Jalili, M. Perc *Information cascades in complex networks*, Journal of Complex Networks, 2017

[18] S. Fortunato, *Community detection in graphs*, Physics Reports 486, 2010

Acknowledgements

First of all, I would like to thank my supervisor at the Central European University, prof. János Kertész, for all the time and the patience he has devoted to me. He has introduced myself in the wonderful world of network science and has always guided and motivated me with enthusiasm and great humanity during these months.

Thanks also to prof. János Török, who, despite the little time spent together, has always been available to answer my doubts and has kindly provided me with useful advice.

Thanks to Mrs. Olga Peredi for the great help in dealing with all the bureaucratic stuff.

Vorrei inoltre rivolgermi in italiano per esprimere alcuni sentiti ringraziamenti.

Innanzitutto, vorrei ringraziare il prof. Andrea Pagnani per la grande disponibilità e la cortesia che mi ha sempre dimostrato in questi mesi e per le rassicurazioni che mi ha saputo offrire nei momenti di maggior preoccupazione.

Un grande grazie a tutta la famiglia Scarra allargata, per le grandi risate, le animate discussioni, i momenti di allegria e di affetto condivisi in tutti questi anni e per la spensieratezza che mi fanno provare tutti i giorni.

Grazie a tutti i Tosti di Pisa, che hanno sempre alleggerito con allegria il peso dell'università e mi hanno fatto passare tre anni meravigliosi e indimenticabili.

Grazie ad Andre, compagno di viaggi e di vita da tanti anni che nonostante la distanza riesce sempre a esprimere parole di affetto e di sostegno.

Grazie a Matte per la pazienza verso le mie martellanti richieste di aiuto e per il fondamentale supporto tecnico.

Grazie a Matte, Fede e Fede per tutti i consigli e per questi bellissimi mesi passati a Budapest.

Grazie a tutti i compagni di corso con cui ho condiviso da vicino le fatiche e le soddisfazioni di questi due anni.

Grazie a Luca e Giacomo che oltre a essere stati un supporto fondamentale in questi due anni sono due grandissimi amici con cui poter condividere, con leggerezza e in ogni momento, insicurezze, progetti e soddisfazioni e che sono sicuro mi accompagneranno ancora in tante avventure.

Grazie ai miei fratelloni Fra e Peppo, per me le più grandi fonti d'ispirazione da quando sono nato. Le loro raccomandazioni e i loro consigli valgono per me più di quelli di qualunque maestro.

Grazie a Frallina, a cui va sempre il mio primo pensiero e che mi sostiene e mi supporta giorno per giorno. La sua presenza gentile e il suo sorriso spensierato mi rassicurano in ogni momento e in ogni cosa che faccio.

Infine, il grazie più grande va ai miei genitori. Non esprimo spesso (per usare un eufemismo) i miei sentimenti e colgo quindi l'occasione per sottolineare l'importanza della loro presenza e soprattutto per esprimere l'immensa gratitudine che provo verso il loro costante sostegno e il loro affetto incondizionato, che sono stati elementi imprescindibili di tutti i piccoli traguardi raggiunti.