Scientific Research Publishing

Advances in Pure Mathematics, 2018, 8, 380-399
http://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Facebook Dynamics: Modelling and Statistical Testing

José Bavio[1], Melina Guardiola[1], Gonzalo Perera[2]

[1]Departamento de Matemática, Universidad Nacional del Sur, Buenos Aires, Argentina
[2]Universidad de la República, Montevideo, Uruguay
Email: jmbavio@yahoo.com.ar, guardiol@uns.edu.ar, pereragonzalo@yahoo.fr

How to cite this paper: Bavio, J., Guardiola, M. and Perera, G. (2018) Facebook Dynamics: Modelling and Statistical Testing. *Advances in Pure Mathematics*, **8**, 380-399.
https://doi.org/10.4236/apm.2018.84021

Received: December 26, 2017
Accepted: April 15, 2018
Published: April 18, 2018

Abstract

In this work we study virtual social networks known as Facebook. It is used by millions of people worldwide, gathering a combination of virtual elements and real world components. We suggest a probabilistic model to describe the long-term behavior of Facebook. This model includes different friendship connection between profiles, directly or by suggestion. Due to web's high interactivity level, we simplify the model assuming Markovian dynamic. After the model is established we propose Complete Transversality (CT) communication concept. CT describes people interaction that reflects profile behaviour and leads to estimators that measure this interaction. Then we introduce a weakness version of CT named Segmental Transversality (ST). Within this framework we develop estimators that allow hypothesis testing of CT and ST. And then, in ST context we propose performance measures to address a priori segmentation's quality.

Keywords

Markov Chains, Random Graphs, Social Networks

1. Introduction

Social networks have emerged as a communication tool of unexpected impact. Frequent contact between people through these networks gives rise to virtual relationships developed according to their interests. This study's approach follows [1] which poses emphasis not only in the individual behaviour but in social interaction within the network.

One of today's challenges is to develop accurate tools to identify influential users by sectors and markets, and to understand information flow dynamics. In turn, it is difficult to fully observe a social network; therefore statistical problems

and probabilistic modelling are important issues (see [2] [3] and [4] for surveys on this subject). Besides this topic, social networks also provide examples of situations of unidentifiable or censored data models and this makes them particularly interesting.

The above concerns lead us to study virtual social networks phenomenon and we restricted our analysis to Facebook. For this we develop a model which, although does not describe it in full reality, it is useful as an approach to network's dynamics study. Naturally this model leads to graph theory which relates to work in [5] and [6].

We will use mathematical tools and statistics from stochastic processes field [7] [8] [9], trying to answer questions such as the existence of transversal communication or how to find if a network segmentation proposed is optimal within a certain communication behaviour between segments.

This paper is structured as follows. In Section 2 we address Facebook modelling using tools of Markov chains, we introduce the concept of complete transversality in the communication, and in this context, we try to find the distribution of random functions involved in the model.

Section 3 proposes two statistical hypothesis tests, one to prove network's CT and the other to prove CT between network segments using U-statistics theory and asymptotic convergence theorems presented in [10]. We end this section, defining useful performance index to measure segmentation's quality.

Section 4 is devoted to conclusions and acknowledgment.

Section 5 is an appendix with proofs of the results obtained in the preceding sections.

Further and related works on this subject can be read in one of author's PhD theses [11].

2. Model Description

In this section we propose a model for describe Facebook dynamics.

Consider $t \in \mathbb{N}$. Let $\mathcal{P}_t$ denote the set of all internet users at instant t and $\mathcal{F}_t$ the set of Facebook's profiles at time t. Of course $\mathcal{P}_t \subseteq \mathcal{P}_{t+1}$ and, we'll also suppose that once a profile is created it cannot be eliminated. (Actually a profile can be eliminated, but this is tedious and difficult to do, so we disregard this behaviour.) Then $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$.

We also denote $\mathcal{P}_\infty$ and $\mathcal{F}_\infty$ to the sets $\mathcal{P}_\infty = \bigcup_{t=0}^{\infty} \mathcal{P}_t$ and $\mathcal{F}_\infty = \bigcup_{t=0}^{\infty} \mathcal{F}_t$ respectively.

For each instant t, we will model the network with a random graph where the nodes represent profiles and the edges represent friendship links.

Definition 1. Let $f, g \in \mathcal{F}_t$. We define the random "friendship" function at time t as the function $\alpha_t : \mathcal{F}_t \times \mathcal{F}_t \to \{0,1\}$, such that if $f \neq g$,

$$\alpha_t(f,g) = \begin{cases} 1, & \text{if } f \text{ and } g \text{ are friend satinstant } t, \\ 0, & \text{if } f \text{ and } g \text{ are not friend satinstant } t, \end{cases}$$

and

$$\alpha_t(f,f)=\begin{cases}1, & \text{for all } f\in\mathcal{F}_t,\\ 0, & \text{if } f\in\mathcal{F}_\infty-\mathcal{F}_t.\end{cases}$$

Remark. The "friendship" relationship is symmetric, so function α_t is also symmetric. Then, the random graph determined by α_t is bi-directional and we can define the adjacency matrix as follows.

Definition 2. Let M be the cardinal number of $\mathcal{F}_\infty$ and let $\mathcal{M}_{M\times M}$ be the set of binary symmetric matrices of order M. We define the random "friendship" matrix at time t as the matrix $\mathcal{A}_t\in\mathcal{M}_{M\times M}$ whose elements are the values taken by α_t for each pair of profiles $(f,g)\in\mathcal{F}_\infty\times\mathcal{F}_\infty$.

Due to the high level of interactivity on social networks, is natural to suppose that $\{\mathcal{A}_t\}$ is a Markov chain with states space in $\mathcal{M}_{M\times M}$. Then, given two states $\mathbb{A}$ and $\mathbb{B}$ of $\{\mathcal{A}_t\}$, we denote $p_{t,t+1}^{\mathbb{A},\mathbb{B}}$ to the one step transition probability from $\mathbb{A}$ to $\mathbb{B}$.

Because of $\mathcal{A}_t$ symmetry's, $p_{t,t+1}^{\mathbb{A},\mathbb{B}}$ is determined by the one step transition probability of the lower subdiagonal's elements of $\mathcal{A}_t$. Then, if we say that f precedes f_j if $i<j$ and we write $f_i\prec f_j$ and, given $\mathbb{A}\in\mathcal{M}_{M\times M}$, we denote $SD(\mathbb{A})=\{\mathbb{A}_{i,j}:i>j\}$ to the set of the lower subdiagonal's entries of $\mathbb{A}$, we have that:

$$p_{t,t+1}^{\mathbb{A},\mathbb{B}}=P\left[SD(\mathcal{A}_{t+1})=SD(\mathbb{B})/SD(\mathcal{A}_t)=SD(\mathbb{A})\right]. \tag{1}$$

For (1) calculation's, we describe through events the profile's actions which have impact on the transition. These events will be linked to certain indices that we will construct to measure affinities between profiles.

We will use the following notation. Let A be any set, we will denote:

$$A\times A-D(A\times A)=\{(a,a'):a,a'\in A,a\neq a'\}$$

to the set of distinct pairs of A's elements, and by

$$A\times A\times A-D(A\times A\times A)=\{(a,a',a''):a,a',a''\in A,a\neq a',a\neq a'',a'\neq a''\}$$

to the set of triples of A's elements they are different by pairs.

Definition 3. Let $p,p'\in\mathcal{P}_\infty$, with $p\neq p'$. We define the p "image index" over p' as the function $X:\mathcal{P}_\infty\times\mathcal{P}_\infty-D(\mathcal{P}_\infty\times\mathcal{P}_\infty)\to\mathbb{N}$ such that

$$\begin{cases}X(p,p')>0, & \text{if } p \text{ has positive image of } p',\\ X(p,p')=0, & \text{if } p \text{ is indiferent to } p',\\ X(p,p')<0, & \text{if } p \text{ has negative image of } p'.\end{cases}$$

Remark. We will suppose that the network lies at steady state, this implies that image between profiles has also evolved to a steady state, so the function X does not depends on time t. Besides, function X is not symmetric, non observable and monotonic.

Definition 4. Let $f,g\in\mathcal{F}_\infty$, with $f\neq g$. We define f"image index" over g as the function $Y_t:\mathcal{F}_\infty\times\mathcal{F}_\infty-D(\mathcal{F}_\infty\times\mathcal{F}_\infty)\to\mathbb{R}$, given by

$$Y_t(f,g)=\frac{1}{m*l}\sum_{i=1}^{m}\sum_{j=1}^{l}\Phi\left(X(p_i,p_j')\right)+\varepsilon_t(f,g), \tag{2}$$

with $\{p_1,\cdots,p_m\}$ and $\{p_1',\cdots,p_l'\}$ the users sets that administrate f and g profiles respectively, $\Phi:\mathbb{R}\to\mathbb{R}$ a monotonic regression function, and $\{\varepsilon_t(f,g)\}$ independent and identically distributed random variables with $E(\varepsilon_t)=0$ and $E(\varepsilon_t^2)=\sigma^2>0$, for all pair $(f,g)\in\mathcal{F}_\infty\times\mathcal{F}_\infty$, $f\neq g$.

To triples ordered of users and triples ordered of profiles we define the following indices.

Definition 5. Let $p,p',p''\in\mathcal{P}_\infty$. For the triple ordered of users (p,p',p'') we define an "image index" as the function

$$U:\mathcal{P}_\infty\times\mathcal{P}_\infty\times\mathcal{P}_\infty-D(\mathcal{P}_\infty\times\mathcal{P}_\infty\times\mathcal{P}_\infty)\to\mathbb{R}$$

that assigns to each triple, a real number that represents acceptance level for the action p suggests to p' to be a friend of p''.

Remark. As with X, U doesn't depend on t, is non symmetric and non observable.

Definition 6. Let $f,g,h\in\mathcal{F}_\infty$ distinct by pairs. For the triple ordered of profiles (f,g,h) we define the "index image" as the function

$$W_t:\mathcal{F}_\infty\times\mathcal{F}_\infty\times\mathcal{F}_\infty-D(\mathcal{F}_\infty\times\mathcal{F}_\infty\times\mathcal{F}_\infty)\to\mathbb{R},$$

given by

$$W_t(f,g,h)=\frac{1}{m*k*l}\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{r=1}^{l}\Psi\left(U(p_i,p_j',p_r'')\right)+\eta_t(f,g,h),\tag{3}$$

with $\{p_1,\cdots,p_m\}$, $\{p_1',\cdots,p_k'\}$ and $\{p_l'',\cdots,p_l''\}$ the users sets that administrate f, g and h profiles respectively, $\Psi:\mathbb{R}\to\mathbb{R}$ a monotonic regression function and $\{\eta_t(f,g,h)\}$ independent and identically distributed random variables with $E(\eta_t)=0$ and $E(\eta_t^2)=\tau^2>0$, for every triple $(f,g,h)\in\mathcal{F}_\infty\times\mathcal{F}_\infty\times\mathcal{F}_\infty$, distinct by pairs.

Then, given $f,g,h\in\mathcal{F}_\infty$, we will suppose that there are $\delta_B>0$, $\delta_R>0$ and $\delta_I>0$, such that:

i) "f breaks friendship with g" $\Leftrightarrow$ $\{Y_t(f,g)<-\delta_B\}$.

(The breakdown of friendship may be due to that f take decision to eliminate g or vice versa).

ii) "f successfully requests friendship to g" $\Leftrightarrow$ $\{Y_t(f,g)>\delta_R\}$.

(The request of successful friendship arises when f asks to g for friendship and g accepts to f as his friend).

iii) "f successfully suggests to g, h's friendship" $\Leftrightarrow$ $\{W_t(f,g,h)>\delta_I\}$.

(The suggestion of successful friendship is given when f suggests h to g, g requests h's friendship and h accepts).

Then, if we denote with $I_t(f)$ to the interventions of f regardless its effects in the transitions from one instant to the next from $\mathbb{A}$ to $\mathbb{B}$, these can be described in a disjunct union of the following events:

$$D_t^1(f)=\text{"}f\text{'s breaks"}=\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=0}}\{Y_t(f,g)<-\delta_B\},$$

$$D_t^2(f) = \text{“}f\text{'s nobreaks”} = \bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=1}} \{Y_t(f,g) \geq -\delta_B\},$$

$$D_t^3(f) = \text{“}f\text{'s requests”} = \bigcup_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}} \{Y_t(f,g) > \delta_R\} \cup \bigcup_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}} \{Y_t(f,g) \leq \delta_R\},$$

$D_t^4(f) = \text{“}f\text{'s suggestions”}$

$$= \bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1}} \left(\bigcup_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}} \{W_t(f,g,h) > \delta_I\} \cup \bigcup_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}} \{W_t(f,g,h) \leq \delta_I\} \right), \text{ that is,}$$

$I_t(f) = \bigcup_{i=1}^{4} D_t^i(f)$, and the transition probabilities in one step from $\mathbb{A}$ to $\mathbb{B}$ can be expressed in the following formula:

$$p_{t,t+1}^{\mathbb{A},\mathbb{B}} = P\left(\bigcap_{f\in\mathcal{F}_\infty} I_t(f) \right). \tag{4}$$

Proposition 1. The one step transition probability from $\mathbb{A}$ to $\mathbb{B}$ of the Markov chain $\{\mathcal{A}_t\}$ is given by:

$$p_{t,t+1}^{\mathbb{A},\mathbb{B}} = \prod_{f\in\mathcal{F}_\infty} \left(\sum_{i=1}^{4} P\left(D_t^i(f)\right) - \sum_{i=1}^{3}\sum_{j=i+1}^{4} P\left(D_t^i(f)\right) P\left(D_t^j(f)\right) \right.$$
$$\left. + \sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{k=j+1}^{4} P\left(D_t^i(f)\right) P\left(D_t^j(f)\right) P\left(D_t^k(f)\right) - \prod_{i=1}^{4} P\left(D_t^i(f)\right) \right),$$

with

$$P\left(D_t^1(f)\right) = \prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=0}} P\left(Y_t(f,g) < -\delta_B\right),$$

$$P\left(D_t^2(f)\right) = \prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=1}} P\left(Y_t(f,g) \geq -\delta_B\right),$$

$$P\left(D_t^3(f)\right) = 1 - \prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}} P\left(Y_t(f,g) \leq \delta_R\right) \prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}} P\left(Y_t(f,g) > \delta_R\right)$$

and

$$P\left(D_t^4(f)\right) = \prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1}} \left[1 - \prod_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}} P\left(W_t(f,g,h) \leq \delta_I\right) \prod_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}} P\left(W_t(f,g,h) > \delta_I\right) \right].$$

Complete Transversality

Complete Transversality (CT) in a social network is associated to a certain communication behavior. This behavior implies that relationship probability between two profiles is always the same. No matter who the profiles are.

CT arises from social scientific theories [12] that poses that massification of social networks would bring an horizontality in communication and transversality in connection between people, overpassing social, economic, ethnographic and other differences. Virtual social networks would bring balance and democracy to all people connection.

In terms of the model, this could be reflected in the "image index" X of one user p over another p' and in the "image index" U of the triple ordered of users (p, p', p'') distinct by pairs. So, averages of regression functions $\Phi(X)$ and $\Psi(U)$ takes the same value in (2) and (3). Let's say they are equal to $C_1 \in \mathbb{R}$ for all distinct user pairs and $C_2 \in \mathbb{R}$ for all triples ordered of users distinct by pairs. Under this conditions, "image indices" given by (2) and (3) are reduced to $C_1 + \varepsilon_t(f,g)$ and $C_2 + \eta_t(f,g,h)$.

When network follows this behaviour, the following results can be established.

Theorem 1. (Homogeneity)

Under the CT context, $\{\mathcal{A}_t\}$ is a time-homogeneous Markov chain.

Theorem 2. (Ergodicity)

Under the CT context, the Markov chain $\{\mathcal{A}_t\}$ is ergodic, and when initial distribution is ergodic, the friendship indicator between a pair of profiles $(f,g) \in \mathcal{F}_\infty \times \mathcal{F}_\infty$, with $f \prec g$, denoted by $\alpha_\infty(f,g)$, is a random variable with Bernoulli distribution and parameter p, $0 < p < 1$, with the same distribution and independent of the friendship indicator of any pair of distinct profiles.

From the results exposed in former Theorem we can conclude that, for $\mathbb{A} \in \mathcal{C}$, ergodic distribution under CT is

$$\pi_{\mathbb{A}}(\infty) \cong \prod_{f \in \mathcal{F}_\infty} \prod_{\substack{g \prec f \\ g \in \mathcal{F}_\infty}} p^{\mathbb{A}_{f,g}} \cdot (1-p)^{1-\mathbb{A}_{f,g}}, \ 0 < p < 1.$$

3. Test and Estimation

We are aimed to discuss the CT's validity in social network Facebook. For this, we will propose two statistics based on samples of N profiles and will study their asymptotic distribution under CT when N increases. Besides, we will present the CT tests.

Further, in this section we will introduce a lighter version of CT called Segment Transversality (ST) related to a given segmentation, that allow us to elaborate segmentation quality index.

3.1. Average Communication between Profiles

Let

$$E_N = \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \alpha_\infty(f_i, f_j), \tag{5}$$

this statistic averages proportion of friends who have profiles on the sample and therefore measures "sample's communication average".

Focusing on long term dynamics, we have seen that under CT, the random

variables $\alpha_\infty\left(f_i, f_j\right)$, $i \neq j$ are independent Bernoulli with parameter p. Then, we can verify that $E\left(E_N\right)=p$ and $Var\left(E_N\right)=\frac{2p(1-p)}{N(N-1)}$.

We want to find E_N's asymptotic distribution, so we study asymptotic behaviour of

$$\begin{aligned} & E_N - p \\ & = \frac{1}{C_N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\alpha_\infty\left(f_i, f_j\right)-p\right) \end{aligned} \tag{6}$$

when sample size N grows towards infinity.

If we denote $Y_N^i = \frac{1}{C_N^2}\sum_{j=i+1}^{N}\left(\alpha_\infty\left(f_i, f_j\right)-p\right)$, we have that $E_N - p = \sum_{i=1}^{N-1} Y_N^i$, where $\left\{Y_N^i : i=1,2,\cdots,N-1\right\}$ forms a triangular array.

Theorem 3. If the triangular array $\left\{Y_N^i : i=1,2,\cdots,N-1\right\}$ verifies the following conditions:

i) For each $N \in \mathbb{N}$, $\left\{Y_N^i : i=1,2,\cdots,N-1\right\}$ are independents;

ii) $E\left(Y_N^i\right) \to 0$, when $N \to \infty$;

iii) $s_N^2 = \sum_{i=1}^{N-1} Var\left(Y_N^i\right) < \infty$;

iv) Exists $\delta > 0$ such that

$$E\left[\left(Y_N^i\right)^{2+\delta}\right] < \infty, \text{ for all } N \text{ and for all } i,$$

and Lyapunov conditions is met, that is,

$$L(N,\delta) = \frac{1}{s_N^{2+\delta}}\sum_{i=1}^{N-1} E\left[\left(Y_N^i\right)^{2+\delta}\right] \to 0, \text{ when } N \to \infty, \tag{7}$$

then

$$\frac{1}{s_N}\sum_{i=1}^{N-1} Y_N^i \overset{w}{\Rightarrow} \mathcal{N}(0,1), \text{ when } N \to \infty. \tag{8}$$

Proof 1. Hypothesis i) is trivial because for $i \neq i'$ and N fixed, $\sum_{j=i+1}^{N}\left(\alpha_\infty\left(f_i, f_j\right)-p\right)$ is independent of $\sum_{j=i'+1}^{N}\left(\alpha_\infty\left(f_{i'}, f_j\right)-p\right)$.

Besides, $E\left(Y_N^i\right) = \frac{1}{C_N^2}(N-i)E\left(\alpha_\infty\left(f_i, f_j\right)-p\right) = 0$, and

$$\begin{aligned} s_N^2 &= \sum_{i=1}^{N-1} Var\left(Y_N^i\right) \\ &= \frac{1}{\left(C_N^2\right)^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Var\left(\alpha_\infty\left(f_i, f_j\right)\right) \\ &= \frac{1}{C_N^2}p(1-p) < \infty \end{aligned} \tag{9}$$

Then, the hypothesis ii) and iii) are met.

Let $\delta = 2$. We will see that iv) holds. For this, we will calculate

$$\begin{aligned} E\left(\left(Y_N^i\right)^4\right) &= \frac{1}{\left(C_N^2\right)^4} E\left[\left(\sum_{j=i+1}^{N}\left(\alpha_\infty\left(f_i, f_j\right)-p\right)\right)^4\right] \\ &= \frac{1}{\left(C_N^2\right)^4} E\left[\sum_{j=i+1}^{N} E\left(\alpha_\infty\left(f_i, f_j\right)-p\right)^4\right. \\ &\quad \left. + C_4^2 \sum_{j=i+1}^{N-1}\sum_{k=j+1}^{N} E\left(\alpha_\infty\left(f_i, f_j\right)-p\right)^2 E\left(\alpha_\infty\left(f_i, f_k\right)-p\right)^2\right] \\ &= \frac{1}{\left(C_N^2\right)^4}\left[\left(N-i\right)\mathbb{C}_p + \left(N-i-1\right)\tilde{\mathbb{C}}_p\right], \end{aligned}$$

with $\mathbb{C}_p = -3p^4 + 3p^3 - 4p^2 + p < 3p^3 + p = \mathbb{C}'_p$, $\tilde{\mathbb{C}}_p = 3p^2\left(1-p\right)^2$, $N-i \le N-1$ and $N-i-1 \le N-2$.

Then,

$$\begin{aligned} E\left(\left(Y_N^i\right)^4\right) &\le \frac{1}{\left(C_N^2\right)^4}\left[\left(N-1\right)\mathbb{C}'_p + \left(N-1\right)\left(N-2\right)\tilde{\mathbb{C}}_p\right] \\ &= \frac{1}{\left(C_N^2\right)^4}\left\{N^2\tilde{\mathbb{C}}_p - \left[\left(3N-2\right)\tilde{\mathbb{C}}_p - \left(N-1\right)\mathbb{C}'_p\right]\right\} \\ &\le \frac{1}{\left(C_N^2\right)^4} N^2\tilde{\mathbb{C}}_p < \infty, \text{ for all } N \text{ and for all } i, \end{aligned}$$

and Lyapunov condition given by (7) can be verified for $\delta = 2$:

$$\begin{aligned} L\left(N,2\right) &= \frac{1}{s_N^4}\sum_{i=1}^{N-1} E\left(\left(Y_N^i\right)^4\right) \\ &\le \frac{1}{s_N^4}\frac{1}{\left(C_N^2\right)^4}\left(N-1\right)N^2\tilde{\mathbb{C}}_p \\ &= \frac{N^2\left(N-1\right)}{\left(C_N^2\right)^2}\frac{\tilde{\mathbb{C}}_p}{p^2\left(1-p\right)^2} \\ &\cong \frac{1}{N}\frac{4\tilde{\mathbb{C}}_p}{p^2\left(1-p\right)^2} \downarrow 0^+, \text{ when } N \to \infty. \end{aligned}$$

Consequently the hypothesis i)-iv) holds and we conclude that

$$\frac{1}{s_N}\sum_{i=1}^{N-1} Y_N^i \overset{w}{\Rightarrow} \mathcal{N}\left(0,1\right), \text{ when } N \to \infty.$$

Returning to the centered statistic expression $E_N - p = \sum_{i=1}^{N} Y_N^i$ and to the expression of s_N given by (9) results the following:

$$\frac{1}{s_N}\sum_{i=1}^{N} Y_N^i \cong \frac{N}{\sqrt{2p\left(1-p\right)}}\left(E_N - p\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0,1\right), \text{ when } N \to \infty,$$

that is,

$$N\left(E_N - p\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0, 2p\left(1-p\right)\right), \text{ when } N \to \infty. \tag{10}$$

As E_N is a communication estimation between profiles, if we select different profile samples under CT, we shouldn't detect differences among p's estimations.

To study this, we propose the following hypothesis test to compare communication average.

Given $f_1, \cdots, f_N$ and $g_1, \cdots, g_N$, independent profile samples of $\mathcal{F}_\infty$, that verifies $\{f_1, \cdots, f_N\} \cap \{g_1, \cdots, g_N\} = \varnothing$, we build the statistics:

$$E_N = \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \alpha_\infty\left(f_i, f_j\right) \text{ and } E_N^* = \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \alpha_\infty\left(g_i, g_j\right)$$

and we have that

$$\mu = E\left(E_N\right) = E\left(E_N^*\right) = p \in (0,1),$$

and

$$\sigma^2 = Var\left(E_N\right) = Var\left(E_N^*\right) = 2p\left(1-p\right),$$

with $\sigma = \sigma(\mu)$, this is, mean and variance are linked and $\sigma(\mu)$ is derivable. So, if we take,

$$g(\mu) = \sqrt{2}\, arcsen\left(\sqrt{\mu}\right),$$

$$g'(\mu) = \frac{1}{\sigma(\mu)} \text{ and,}$$

$$N\left(g\left(E_N\right) - g\left(p\right)\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0, \sigma^2\left(p\right) g'\left(p\right)^2\right) = \mathcal{N}(0,1), \quad \text{when } N \to \infty, \tag{11}$$

and

$$N\left(g\left(E_N^*\right) - g\left(p\right)\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0, \sigma^2\left(p\right) g'\left(p\right)^2\right) = \mathcal{N}(0,1), \quad \text{when } N \to \infty. \tag{12}$$

Consequently, to test average communication, we can substract (11) and (12). This statistic, under CT assumption, results

$$N\left(g\left(E_N\right) - g\left(E_N^*\right)\right) \overset{w}{\Rightarrow} \mathcal{N}(0,2), \quad \text{when } N \to \infty. \tag{13}$$

and, given a signification level α, we obtain the following critic region

$$R_\alpha = \left\{ \frac{N}{\sqrt{2}} \left| g\left(E_N\right) - g\left(E_N^*\right) \right| \geq z_{\alpha/2} \right\}.$$

Real Data Testing

We perform such an experiment in a real profile network that gives permission to the authors for sampling. For confidential reasons we cannot release any of the data used to make calculations. We can state that we take two independent and disjunct samples of size $N = 75$. The statistic value was $\frac{N}{\sqrt{2}} \left| g\left(E_N\right) - g\left(E_N^*\right) \right| = 2.1$. Therefore, with a $\alpha = 0.05$ significance level, we

reject *CT* hypothesis.

This results indicates the social network Facebook is a platform in which communication between people or groups of people is it NOT TRANSVERSAL.

3.2. Mean Square Deviation of the Communication between Profiles

Let

$$\begin{aligned} T_N &= \frac{1}{C_N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left[\alpha_\infty(f_i,f_j)-\frac{1}{C_N^2}\sum_{k=1}^{N-1}\sum_{l=k+1}^{N}\alpha_\infty(f_k,f_l)\right]^2 \\ &= \frac{1}{C_N^2}\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\alpha_\infty(f_i,f_j)^2-\frac{1}{C_N^2}\left(\sum_{k=1}^{N-1}\sum_{l=k+1}^{N}\alpha_\infty(f_k,f_l)\right)^2\right] \\ &= \frac{1}{\left(C_N^2\right)^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{k=1}^{N-1}\sum_{l=k+1}^{N}\frac{\left(\alpha_\infty(f_i,f_j)-\alpha_\infty(f_k,f_l)\right)^2}{2}, \end{aligned} \tag{14}$$

be an statistic that estimates the mean square deviation of the communication between profiles respect to its mean.

In order to find the asymptotic distribution of T_N we use properties of U-statistics introduced in [13].

Suppose that $X_1,\cdots,X_N$ are independent and identically distributed random variables and that $h:\mathbb{R}^r\to\mathbb{R}$, $1\le r\le N$ is some symmetric function respect to permutations.

Definition 7. A U-statistic of order r with kernel h is defined as

$$U=U_N=\frac{1}{C_N^r}\sum_{1\le i_1<i_2<\cdots<i_r\le N}h\left(X_{i_1},\cdots,X_{i_r}\right).$$

We state the following theorem whose proof can be seen in [10].

Theorem 4. Let U_N be a U-statistic of order r with kernel h. Suppose that $E\left(h^2\left(X_1,\cdots,X_r\right)\right)<\infty$ and that

i) $X_1,X_2,\cdots,X_r,Y_2,\cdots,Y_r$ are independent and identically distributed random variables,

ii) $\theta=E\left(h\left(X_1,\cdots,X_r\right)\right)$ and,

iii) $\zeta_1=Cov\left(h\left(X_1,X_2,\cdots,X_r\right),h\left(X_1,Y_2,\cdots,Y_r\right)\right)$.

Then, if $0<\zeta_1<\infty$,

$$\sqrt{N}\left(U_N-\theta\right)\overset{w}{\Rightarrow}\mathcal{N}\left(0,r^2\zeta_1\right),\quad \text{when } N\to\infty. \tag{15}$$

Then we have that T_N is a U-statistic of order 2 with kernel

$$h\left(v_\infty(f_i),v_\infty(f_k)\right)=\frac{\left(\alpha_\infty(f_i,f_j)-\alpha_\infty(f_k,f_l)\right)^2}{2},$$

where $v_\infty(f_i)=\left(\alpha_\infty(f_i,f_h)\right)_{h\in\mathcal{F}_\infty}$ and $v_\infty(f_k)=\left(\alpha_\infty(f_k,f_h)\right)_{h\in\mathcal{F}_\infty}$. Thus, under *CT*, $\theta=p(1-p)$ and $\zeta_1=\frac{1}{4}p(1-p)-p^2(1-p)^2$.

Remark. For the proposed model for Facebook, p is the friendship probability

of any pair of profiles in long term and, because of the large size of the network, p is probably less than 0.5. Thus, $p \neq 0$, $p \neq 1$ and $p \neq 0.5$, therefore $\zeta_1 \neq 0$.

As $0 < \zeta_1 < \infty$, by the Theorem 4 is

$$\sqrt{N}\left(T_N - p(1-p)\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0, p(1-p) - 4p^2(1-p)^2\right), \quad \text{when } N \to \infty. \quad (16)$$

As $\mu = p(1-p)$ and $\sigma = \sqrt{p(1-p) - 4p^2(1-p)^2}$, we see that $\sigma = \sigma(\mu)$ and $\sigma(\mu)$ is derivable. Taking

$$g(\mu) = arcsen\left(2\sqrt{\mu}\right),$$

$g(\mu)$ verifies that $g'(\mu) = \dfrac{1}{\sigma(\mu)}$ and the limit expression on (16) is

$$\sqrt{N}\left(g(T_N) - g\left(p(1-p)\right)\right) \overset{w}{\Rightarrow} \mathcal{N}\left(0, \sigma^2\left(p(1-p)\right) g'\left(p(1-p)\right)^2\right) = \mathcal{N}(0,1), \quad (17)$$

when $N \to \infty$.

We can make a test to prove CT by comparing the mean square deviation of two independent populations of profiles. For this we take two independent samples of N profiles of $\mathcal{F}_\infty$, $f_1, \cdots, f_N$ and $g_1, \cdots, g_N$ such that $\{f_1, \cdots, f_N\} \cap \{g_1, \cdots, g_N\} = \varnothing$, we construct the statistics

$$T_N = \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\alpha_\infty(f_i, f_j) - \frac{1}{C_N^2} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \alpha_\infty(f_k, f_l) \right]^2$$

and

$$T_N^* = \frac{1}{C_N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\alpha_\infty(g_i, g_j) - \frac{1}{C_N^2} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \alpha_\infty(g_k, g_l) \right]^2,$$

then, by (17), we obtain the asymptotic distribution of the statistic for the comparison of mean square deviations under the context of CT, that is,

$$\sqrt{N}\left(g(T_N) - g\left(T_N^*\right)\right) \overset{w}{\Rightarrow} \mathcal{N}(0,2), \; N \to \infty,$$

being the critical region at the level of significance α

$$R_\alpha = \left\{ \sqrt{\frac{N}{2}} \left| g(T_N) - g\left(T_N^*\right) \right| \geq z_{\alpha/2} \right\}.$$

Similarly as for the test comparing the average proportion of communication between profiles, we use the same real profile and, on his network of friends, we take two independent and disjoint samples and calculate the statistic and the critical region, concluding that the hypothesis of CT is rejected again.

3.3. Segmented Transversality

A context of CT in a social network like Facebook is far from reality as evidenced by the findings of the two tests that we made. Is reasonable to think that profiles tend to cluster in different segments according to social criteria such as political ideologies, economic interests, musical tastes, ages, etc., and that these

segments are also related to each other.

We introduce the concept of Segmented Transversality (*ST*), that is, *CT* between segments. Then, making a priori segmentation on the network, we will introduce a statistic representing the communication between pairs of segments and we will prove *CT* between the profiles of the segments.

Let $S_1,\cdots,S_k$ be a partition in segments of $\mathcal{F}_\infty$. We notice with $a_i=\frac{card\{S_i\}}{card\{\mathcal{F}_\infty\}}$ to the proportion of profiles of S_i, $i=1,2,\cdots,k$, $a_i>0$, $\sum_{i=1}^{k}a_i=1$, and we make a random stratified sampling by segments as follows: $f_1,\cdots,f_{[a_1N]}$ are chosen randomly inside of S_1, $f_{[a_1N]+1},\cdots,f_{[(a_1+a_2)N]}$ are chosen randomly inside of S_2, and so on until $f_{\left[\left(\sum_{i=1}^{k-1}a_i\right)N\right]+1},\cdots,f_{\left(\sum_{i=1}^{k}a_i\right)N}$ are chosen randomly inside of S_k, being $[x]$ the integer part of x, that is, $[x]=\max\{k\in\mathbb{N}:k\le x\}$, for $x>0$.

Given the sets $I_1=\{1,\cdots,[a_1N]\}$, $I_2=\{[a_1N]+1,\cdots,[(a_1+a_2)N]\}$, $\cdots$, $I_k=\left\{\left[\left(\sum_{i=1}^{k-1}a_i\right)N\right]+1,\cdots,\left(\sum_{i=1}^{k}a_i\right)N\right\}$ and, given S_r and S_t two segments of the partition on k segments of $\mathcal{F}_\infty$, for $f_i\in S_r$ and $f_j\in S_t$, with $i\in I_r$ and $j\in I_t$, we notice with q_{ij} to the probability of friendship between f_i and f_j, that is, $q_{ij}:=P\left(\alpha_\infty\left(f_i,f_j\right)=1\right)=q_{ji}$.

Remark. Friendship's random functions $\left(\alpha_\infty\left(f_i,f_j\right)\right)_{i<j}$ still are independent random variables with Bernoulli distribution, but now the parameter distribution depends on the intensity of the relationship between the pair of profiles considered.

Definition 8. Given two segments of profiles S_r and S_t, we say that there is *CT* between them if given $f_i\in S_r$ and $f_j\in S_t$, with $i\in I_r$ and $j\in I_t$, $\alpha_\infty\left(f_i,f_j\right)$ has Bernoulli distribution with parameter q_{rt}, for all $f_i\in S_r$ and for all $f_j\in S_t$.

That is, *CT* between segments means a distinctive homogeneous behavior in the communication between them.

So, let

$$E(r,t)=\frac{1}{card\{I_r\}}\sum_{i\in I_r}\frac{1}{card\{I_t\}}\sum_{j\in I_t}\alpha_\infty\left(f_i,f_j\right),\tag{18}$$

be the average proportion of friends of the profiles of S_r in the segment S_t. Then, under *CT* between S_r and S_t, we have that $E\left(E(r,t)\right)=q_{rt}$ and $Var\left(E(r,t)\right)=\frac{q_{rt}(1-q_{rt})}{card\{I_r\}card\{I_t\}}$.

Of the same way as we obtain the asymptotic distribution of the centered estimator E_N-p of the expression (6), we can obtain the asymptotic distribution of

$$E(r,t)-q_{rt}=\sum_{i\in I_r}\frac{1}{card\{I_r\}}\sum_{j\in I_t}\frac{\alpha_\infty(f_i,f_j)-q_{rt}}{card\{I_t\}}, \tag{19}$$

resulting

$$\sqrt{card\{I_r\}card\{I_t\}}\left(E(r,t)-q_{rt}\right)\overset{w}{\Rightarrow}\mathcal{N}\left(0,q_{rt}(1-q_{rt})\right), \quad \text{when } N\to\infty. \tag{20}$$

If we want to test CT between a pair of segments S_r and S_t we make a stratified random sampling independent from the previous one in which $g_1,\cdots,g_{[a_1N]}$ are chosen randomly inside of S_1, $g_{[a_1N]+1},\cdots,g_{\lceil(a_1+a_2)N\rceil}$ are chosen randomly inside of S_2, and so on until $g_{\left\lceil\left(\sum_{i=1}^{k-1}a_i\right)N\right\rceil+1},\cdots,g_{\left(\sum_{i=1}^{k}a_i\right)N}$ are chosen randomly inside of S_k, with

$$\left\{f_1,\cdots,f_{[a_1N]}\right\}\cap\left\{g_1,\cdots,g_{[a_1N]}\right\}=\varnothing,$$

$$\left\{f_{[a_1N]+1},\cdots,f_{\lceil(a_1+a_2)N\rceil}\right\}\cap\left\{g_{[a_1N]+1},\cdots,g_{\lceil(a_1+a_2)N\rceil}\right\}=\varnothing,$$

$$\left\{f_{\left\lceil\left(\sum_{i=1}^{k-1}a_i\right)N\right\rceil+1},\cdots,f_{\left(\sum_{i=1}^{k}a_i\right)N}\right\}\cap\left\{g_{\left\lceil\left(\sum_{i=1}^{k-1}a_i\right)N\right\rceil+1},\cdots,g_{\left(\sum_{i=1}^{k}a_i\right)N}\right\}=\varnothing,$$

and we construct the statistic

$$E^*(r,t)=\frac{1}{card\{I_r\}}\sum_{i\in I_r}\frac{1}{card\{I_t\}}\sum_{j\in I_t}\alpha_\infty(g_i,g_j). \tag{21}$$

Then, under CT between S_r and S_t, are $E\left(E^*(r,t)\right)=q_{rt}$, $Var\left(E^*(r,t)\right)=\frac{q_{rt}(1-q_{rt})}{card\{I_r\}card\{I_t\}}$ and

$$\sqrt{card\{I_r\}card\{I_t\}}\left(E^*(r,t)-q_{rt}\right)\overset{w}{\Rightarrow}\mathcal{N}\left(0,q_{rt}(1-q_{rt})\right), \quad \text{when } N\to\infty. \tag{22}$$

For $E(r,t)$ and $E^*(r,t)$ we have that the variance σ is a function of the expected value μ, that is, $\sigma(\mu)=\mu(1-\mu)$, where $\mu=q_{rt}$ and $\sigma(\mu)$ is derivable. Then, taking

$$g(\mu)=2arcsen\left(\sqrt{\mu}\right)$$

we have that $g'(\mu)=\frac{1}{\sigma(\mu)}$ and, similarly as in the previous section, we can conclude that

$$\sqrt{card\{I_r\}card\{I_t\}}\left(g\left(E(r,t)\right)-g\left(E^*(r,t)\right)\right)\overset{w}{\Rightarrow}\mathcal{N}(0,2), \quad \text{when } N\to\infty,$$

being the critical region at the level of significance α,

$$R_\alpha=\left\{\sqrt{\frac{card\{I_r\}card\{I_t\}}{2}}\left|g\left(E(r,t)\right)-g\left(E^*(r,t)\right)\right|\geq z_{\frac{\alpha}{2}}\right\}.$$

Therefore, if the test leads to the rejection of the hypothesis of CT between the segments S_r and S_t, with an error probability of α we say that such segments

do not have a distinctive homogeneous behavior in the communication.

3.4. Quality on Segmentation

If we divide the network into k disjoint segments, we can take all possible pairs of those k segments, C_k^2, and make a total of C_k^2 test, one of each pair and test whether the segment of this pair have a distinctive homogeneous behavior in the communication. We can represent these C_k^2 test by a binary symmetric matrix of order k, $\mathcal{S}$, in which each element $\mathcal{S}_{ij}$ is one if the segments S_i and S_j were not homogeneous in terms of communication, that is, if the corresponding test rejects the hypothesis of CT and, $\mathcal{S}_{ij}$ equals zero, otherwise.

Then, noticing the cardinal of the set of "ones" in the subdiagonal of $\mathcal{S}$ as

$$g(d) = card\{1's \text{ in } SD(\mathcal{S})\},$$

we can define the following useful performance index to measure the quality on segmentation:

$$C_p = \frac{g(d)}{C_k^2} 100\%. \tag{23}$$

If we keep the segmentation and make a stratified random sampling segment m times, with m sufficiently large, and calculate m times the index defined in (23), C_p^i, with $i = 1, \cdots, m$, we can observe the histogram representing the distribution of quality on segmentation. If the most of the times this measure results, for example, greater than the mean of the observations, we continue segmenting according to the criteria which has been used, otherwise it is desirable to modify the segmentation criteria.

Let's illustrate this with an example. Suppose we segment people by age (>15, <20, >20 and <30, >30) and gender (M, F).

Age\Gender	M	F
>15, <20	S_1	S_2
>20, <30	S_3	S_4
>30	S_5	S_6

Given this segmentation, suppose we conduct the 15 hypothesis test for segment transversality and obtain the following segment adjacency matrix

Segments	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0					
S_2	0	0				
S_3	1	0	0			
S_4	1	1	0	0		
S_5	0	0	1	0	0	
S_6	1	0	1	0	0	1

A one in this matrix means that segment S_i with segment S_j behaves distinctly in the sense of transversality communication. Zeros in the upper triangle matrix doesn't mean anything. Each segment compare to each self is homegenous (zero in the diagonals) except in segment 6. This could mean that further segmentation within it might improve audience segmentation. Nevertheless, segment 6 distinguish from other segments anyway.

If we sum the ones under the diagonal and divided by the number of segment combinations (see (23)), we calculate the performance of this segmentation which is 40%. The more the zeros the lower the performance index.

Following this framework we can improvement this segmentation by:

1) fusion homogenous segments

2) explore intra-segmentation in the cases were there was one in the diagonal

For actions in the first group, we look that in this toy example S_1 and S_2 doesn't differ in communication behaviour, so we could consider to group them as one segment. We could summaries this saying that gender doesn't segment among young people (under 20 years). We calculate again the matrix with this augmented segment and calculate the performance. Then occam's razor lead us to select the least segmented partition when we have two or more with same performance.

For segment 6, this is female over thirty years, we calculate that it is inhomogenous with itself, so we could try a sub-segmentation by education degree or by motherhood. Then repeat the five tests between the new segment with the previous ones and calculate performance of this new matrix with one or more rows.

Of course this iterative method implies significant work with estimation, data recompilation, amount of data, independent sampling, etc. that although might be of extreme relevance it's beyond the scope of pure mathematics and poses a great source of scientific challenge and interdisciplinary work.

4. Conclusions

In this work we analyze Facebook social network, modelling it with a Markov Chain and several random variables representing profiles friendships. We further propose communication behaviour between all profiles called complete transversality that assumes no bias between profiles willingness to connect as friends. This CT behaviour leads to estimations that allow us a hypothesis test by means of mean square deviation to reject the CT assumption. This might be an obvious conclusion (because people behaves within Facebook as they behave in real context), but it has all the hypothesis testing machinery behind which gives it strong rigorosity.

Next step in our work was to weaken CT and, for this, we introduce ST (segment Transversality). This is given a determined network segmentation, and each segment profile connects to any other segment profile with the same probability. (Of course this probability can change with different pair of segments.)

In this ST scenario we were able to compare between two entire segments and

determine whether they behave in the same way or differently. If we don't find differences we can safely group the two segments in order to improve the original segmentation. For a given segmentation we test intra segment and every pair of different segments and define a performance index that reflects a percentage of the segments that behaves differently from all the possible comparisons. With this result, we join similar segment (don't reject null hypothesis) and recalculate tests and performance index to the new segmentation.

This iterative process of grouping homogenous segments or leaving different leads to a hierarchy in segmentations according to segmentation's quality. This information could be interpreted and used to determine if segmentation's rules are adequate to distinguish segments; of course the comparison falls in this communication behavior sense.

References

[1] Freeman, L.C. (2006) The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press, Vancouver.

[2] Rao, A.R., Bandyopadhyay, S. and Sinha, B.K. (2011) Models of Social Networks with Statistical Applications. SAGE Publications Ltd., Thousand Oaks.

[3] Furht, B. (2010) Handbook of Social Network Technologies and Applications. Springer, Berlin. https://doi.org/10.1007/978-1-4419-7142-5

[4] Scott, J., Carrington, P. and Wasserman, S. (2005) Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge.

[5] Ralescu, D.A., Chen, L. and Peng, J. (2017) Uncertain Vertex Coloring Problem. *Soft Computing*, **21**, 1-10.

[6] Rosyida, I., Peng, J., Chen, L., Widodo, W., Indrati, Ch.R. and Sugeng, K.A. (2016) An Uncertain Chromatic Number of an Uncertain Graph Based on Alpha-Cut Coloring. *Fuzzy Optimization and Decision Making*, **17**, 103-123.

[7] Meyn, S.P. and Tweedie, R.L. (1993) Markov Chains and Stochastic Stability. Springer, New York. https://doi.org/10.1007/978-1-4471-3267-7

[8] De Meer, H., Bolch, G., Greiner, S. and Trivedi, K. (2006) Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications. Wiley and Sons, Hoboken.

[9] DasGupta, A. (2008) Asymptotic Theory of Statistics and Probability. Springer, Berlin.

[10] Sering, R. (1980) Approximation Theorems of Mathematical Statistics. John Wiley and Sons, New York, Chichester, Brisbane, Toronto. https://doi.org/10.1002/9780470316481

[11] Guardiola, M. (2013) Estadística de procesos estocásticos y aplicaciones a las redes sociales. http://repositoriodigital.uns.edu.ar/bitstream/123456789/446/1/TESIS%20DOCTORAL%20Melina%20Guardiola.pdf

[12] De Ugarte, D. (2011) El poder de las redes. Manual ilustrado para ciberactivistas. Colección Biblioteca de las Indias Electrónicas.

[13] Hoeffding, W. (1948) A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics*, **19**, 293-325. https://doi.org/10.1214/aoms/1177730196

Appendix

1.1. Proof of Proposition 1

Let $f' \in \mathcal{F}_\infty$, $f' \neq f$ and $I_t(f) = \bigcup_{i=1}^{4} D_t^i(f)$. We will see that $I_t(f')$ is independent of $I_t(f)$. For this, it is enough prove that $D_t^i(f')$ is independent from $D_t^i(f)$, when $i = 1,2,3,4$. In fact, if we look at $D_t^i(f')$, the first coordinate of both involved indices $Y_t(f',g)$ for $i = 1,2,3$ and $W_t(f',g,h)$ for $i = 4$, is fixed and is the first coordinate of the random variables $\varepsilon_t(f',g)$ and $\eta_t(f',g,h)$ of the mentioned indices respectively. Then, $Y_t(f',g)$ is independent from $Y_t(f,g)$ and $W_t(f',g,h)$ is independent from $W_t(f,g,h)$, hence, $D_t^i(f')$ is independent from $D_t^i(f)$, when $i = 1,2,3,4$. Thus,

$$\begin{aligned} p_{t,t+1}^{\mathbb{A},\mathbb{B}} &= P\left(\bigcap_{f \in \mathcal{F}_\infty} I_t(f)\right) \\ &= \prod_{f \in \mathcal{F}_\infty} P\left(I_t(f)\right). \end{aligned}$$

Applying inclusion-exclusion principle for the events $D_t^i(f)$, $i = 1,2,3,4$, we have

$$\begin{aligned} p_{t,t+1}^{\mathbb{A},\mathbb{B}} = \prod_{f \in \mathcal{F}_\infty} \Bigg(& \sum_{i=1}^{4} P\left(D_t^i(f)\right) - \sum_{i=1}^{3} \sum_{j=i+1}^{4} P\left(D_t^i(f) \cap D_t^j(f)\right) \\ & + \sum_{i=1}^{2} \sum_{j=i+1}^{3} \sum_{k=j+1}^{4} P\left(D_t^i(f) \cap D_t^j(f) \cap D_t^k(f)\right) - P\left(\bigcap_{i=1}^{4} D_t^i(f)\right)\Bigg) \end{aligned} \tag{24}$$

We have to check that the events $D_t^i(f)$, $i = 1,2,3,4$, are independent between them for a same profile f. $D_t^4(f)$ is independent from $D_t^1(f)$, from $D_t^2(f)$ and $D_t^3(f)$, because $\eta_t(f,g,h)$, which corresponds to image index $W_t(f,g,h)$ involved in $D_t^4(f)$ is independent from $\varepsilon_t(f,g)$ which corresponds to index $Y_t(f,g)$ that appears in $D_t^1(f)$, $D_t^2(f)$ and $D_t^3(f)$.

$D_t^1(f)$ is independent from $D_t^2(f)$ and from $D_t^3(f)$, because if a profile $f \in \mathcal{F}_\infty$ appears in the intersections of $D_t^1(f)$, it doesn't figure either in the intersections of $D_t^2(f)$ nor in the unions of $D_t^3(f)$. The same argument allows us to affirm that $D_t^2(f)$ is independent from $D_t^3(f)$.

Therefore, transition probability (24) is given by

$$\begin{aligned} p_{t,t+1}^{\mathbb{A},\mathbb{B}} = \prod_{f \in \mathcal{F}_\infty} \Bigg(& \sum_{i=1}^{4} P\left(D_t^i(f)\right) - \sum_{i=1}^{4} \sum_{j=i+1}^{4} P\left(D_t^i(f)\right) P\left(D_t^j(f)\right) \\ & + \sum_{i=1}^{4} \sum_{j=i+1}^{4} \sum_{k=j+1}^{4} P\left(D_t^i(f)\right) P\left(D_t^j(f)\right) P\left(D_t^k(f)\right) - \prod_{i=1}^{4} P\left(D_t^i(f)\right)\Bigg). \end{aligned}$$

We will see how $P\left(D_t^i(f)\right)$, $i = 1,2,3,4$ is finally written.

If $g, g' \in \mathcal{F}_\infty$, with $g \prec f$, $g' \prec f$, $g \neq g'$, $\varepsilon_t(f,g)$ is independent from $\varepsilon_t(f,g')$. Then, are independents: $\{Y_t(f,g) < -\delta_B\}$ from $\{Y_t(f,g') < -\delta_B\}$, $\{Y_t(f,g) \geq -\delta_B\}$ from $\{Y_t(f,g') \geq -\delta_B\}$, $\{Y_t(f,g) \leq \delta_R\}$ from $\{Y_t(f,g') \leq \delta_R\}$ and $\{Y_t(f,g) > \delta_R\}$ from $\{Y_t(f,g') > \delta_R\}$. Thus,

$$P\left(D_t^1(f)\right)=P\left(\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=0}}\left\{Y_t(f,g)<-\delta_B\right\}\right)=\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=0}}P\left(Y_t(f,g)<-\delta_B\right),$$

$$P\left(D_t^2(f)\right)=P\left(\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=1}}\left\{Y_t(f,g)\geq-\delta_B\right\}\right)=\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=1}}P\left(Y_t(f,g)\geq-\delta_B\right),$$

and, applying set properties we have that

$$\begin{aligned}P\left(D_t^3(f)\right)&=P\left(\bigcup_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}}\left\{Y_t(f,g)>\delta_R\right\}\cup\bigcup_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}}\left\{Y_t(f,g)\leq\delta_R\right\}\right)\\&=1-P\left(\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}}\left\{Y_t(f,g)\leq\delta_R\right\}\cap\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}}\left\{Y_t(f,g)>\delta_R\right\}\right)\\&=1-\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}}P\left(Y_t(f,g)\leq\delta_R\right)\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}}P\left(Y_t(f,g)>\delta_R\right),\end{aligned}$$

as if $g'\in\left\{g\prec f:\mathbb{A}_{f,g}=0,\mathbb{B}_{f,g}=1\right\}$ then $g'\notin\left\{g\prec f:\mathbb{A}_{f,g}=0,\mathbb{B}_{f,g}=0\right\}$ and accordingly we have that $\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}}\left\{Y_t(f,g)\leq\delta_R\right\}$ and $\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}}\left\{Y_t(f,g)>\delta_R\right\}$ are independents.

Finally,

$$\begin{aligned}P\left(D_t^4(f)\right)&=P\left(\bigcap_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1}}\left(\bigcup_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}}\left\{W_t(f,g,h)>\delta_I\right\}\cup\bigcup_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}}\left\{W_t(f,g,h)\leq\delta_I\right\}\right)\right)\\&=\prod_{\substack{g\prec f:\\ \mathbb{B}_{f,g}=1}}\left(1-P\left(\bigcap_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}}\left\{W_t(f,g,h)\leq\delta_I\right\}\cap\bigcap_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}}\left\{W_t(f,g,h)>\delta_I\right\}\right)\right),\end{aligned}$$

and, for fixed $f,g\in\mathcal{F}_\infty$, $h,h'\in\mathcal{F}_\infty$, $h\neq h'$, $h\prec f$, $h'\prec f$, $\eta_t(f,g,h)$ is independent from $\eta_t(f,g,h')$. So, are independents $\left\{W_t(f,g,h)\leq\delta_I\right\}$ from $\left\{W_t(f,g,h')\leq\delta_I\right\}$ and $\left\{W_t(f,g,h)>\delta_I\right\}$ from $\left\{W_t(f,g,h')>\delta_I\right\}$. Besides, $\bigcap_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}}\left\{W_t(f,g,h)\leq\delta_I\right\}$ and $\bigcap_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}}\left\{W_t(f,g,h)>\delta_I\right\}$ are independents, as if $h'\in\left\{h\prec f:\mathbb{A}_{f,h}=1,\mathbb{B}_{g,h}=1\right\}$ then $h'\notin\left\{h\prec f:\mathbb{A}_{f,h}=1,\mathbb{B}_{g,h}=0\right\}$. Therefore,

$$P\left(D_t^4(f)\right)=\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1}}\left(1-\prod_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}}P\left(\left\{W_t(f,g,h)\le\delta_I\right\}\right)\prod_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}}P\left(\left\{W_t(f,g,h)>\delta_I\right\}\right)\right).$$

1.2. Proof of Theorem 1

Suppose we are in a *CT* context. To prove time homogeneity in Markov chain $\{\mathcal{A}_t\}$ we will see that probabilities involved in $p_{t,t+1}^{\mathbb{A},\mathbb{B}}$, $P\left(D_t^i(f)\right)$, $i=1,2,3,4$, don't depend on time. To do this we introduce some useful notations.

Let F and G be the distribution function for random variables $\varepsilon_t(f,g)$ and $\eta_t(f,g,h)$ related with image indices $Y_t(f,g)$ and $W_t(f,g,h)$ respectively.

The *CT* hypothesis implies that F and G neither depend on time nor on profiles. This is, $P\left(\varepsilon_t(f,g)\le x\right)=F(x)$ and $P\left(\eta_t(f,g,h)\le x\right)=G(x)$, for all $t=0,1,\cdots$, for all $f,g\in\mathcal{F}_\infty$ and every ordered triple $f,g,h\in\mathcal{F}_\infty$, with F and G continuous functions.

We also denote with $c_{\mathbb{A},\mathbb{B}}^{i,j}(f)$ and $d_{\mathbb{A},\mathbb{B}}^{1,j}(f,g)$ the following cardinals numbers of sets:

$$c_{\mathbb{A},\mathbb{B}}^{i,j}(f)=card\left\{g\prec f:\mathbb{A}_{f,g}=i,\mathbb{B}_{f,g}=j;i,j\in\{0,1\}\right\}$$

and

$$d_{\mathbb{A},\mathbb{B}}^{1,j}(f,g)=card\left\{h\prec f:\mathbb{A}_{f,h}=1,\mathbb{B}_{h,g}=j;j\in\{0,1\}\right\}.$$

Consequently, under *CT* hypothesis we have,

$$P\left(D_t^1(f)\right)=\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=0}}P\left(C_1+\varepsilon_t(f,g)<-\delta_B\right)=F\left(-\delta_B-C_1\right)^{c_{\mathbb{A},\mathbb{B}}^{1,0}(f)},$$

$$P\left(D_t^2(f)\right)=\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1,\\ \mathbb{B}_{f,g}=1}}P\left(C_1+\varepsilon_t(f,g)\ge-\delta_B\right)=\left(1-F\left(-\delta_B-C_1\right)\right)^{c_{\mathbb{A},\mathbb{B}}^{1,1}(f)},$$

$$\begin{aligned}P\left(D_t^3(f)\right)&=1-\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=1}}P\left(C_1+\varepsilon_t(f,g)\le\delta_R\right)\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=0,\\ \mathbb{B}_{f,g}=0}}P\left(C_1+\varepsilon_t(f,g)>\delta_R\right)\\&=1-F\left(\delta_R-C_1\right)^{c_{\mathbb{A},\mathbb{B}}^{0,1}(f)}\left(1-F\left(\delta_R-C_1\right)\right)^{c_{\mathbb{A},\mathbb{B}}^{0,0}(f)}\end{aligned}$$

and

$$\begin{aligned}&P\left(D_t^4(f)\right)\\&=\prod_{\substack{g\prec f:\\ \mathbb{A}_{f,g}=1}}\left(1-\prod_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=1}}P\left(C_2+\eta_t(f,g,h)\le\delta_I\right)\prod_{\substack{h\prec f:\\ \mathbb{A}_{f,h}=1,\\ \mathbb{B}_{g,h}=0}}P\left(C_2+\eta_t(f,g,h)>\delta_I\right)\right)\\&=\left(1-G\left(\delta_I-C_2\right)^{d_{\mathbb{A},\mathbb{B}}^{1,1}(f,g)}\left(1-G\left(\delta_I-C_2\right)\right)^{d_{\mathbb{A},\mathbb{B}}^{1,0}(f,g)}\right)^{c_{\mathbb{A},\mathbb{B}}^{1,1}(f)+c_{\mathbb{A},\mathbb{B}}^{1,0}(f)}.\end{aligned}$$

Then, $P\left(D_t^i(f)\right)$, $i=1,2,3,4$, and therefore $p_{t,t+1}^{\mathbb{A},\mathbb{B}}$ doesn't depend on *t*.

Consequently, $\{\mathcal{A}_t\}$ is *homogeneous* in time.

1.3. Proof of Theorem 2

Let $S = \mathcal{M}_{M\times M}$ be the space states of $\{\mathcal{A}_t\}$ and suppose that we are in a *CT* context.

We denote $\mathcal{C} = \{\mathbb{A} \in S : \mathbb{A}_{i,i} = 1, \text{for all } i, i = 1,\cdots,M\}$ to the set of states with all ones on the diagonal.

Lemma 1. $\mathcal{C}$ is a closed, irreducible and aperiodic communication class.

Proof 1. Let $\mathbb{A}' \in S$ such that $\mathbb{A}' \notin \mathcal{C}$. Then, for some $i = 1,\cdots,M$, $\mathbb{A}'_{i,i} = 0$, that is, at time *t*, the friendship state $\mathbb{A}'$ indicates that for some $i = 1,\cdots,M$, the profile f_i doesn't exists on Facebook. As we have supposed that once a profile is created it can't be deleted, then if $\mathbb{A} \in \mathcal{C}$, $p_{t,t+1}^{\mathbb{A},\mathbb{A}'} = 0$, that is, an state outside of $\mathcal{C}$ is not accessible from a state inside of $\mathcal{C}$. Then, $\mathcal{C}$ is closed.

Also, given $\mathbb{A}, \mathbb{B} \in \mathcal{C}$, outside the diagonal, an entry equal to one can turn into to in zero, or an entry with zero can turn into one in one step, that is, $\mathbb{A}$ and $\mathbb{B}$ are communicated and aperiodic, so $\mathcal{C}$ is irreducible and aperiodic.

As we are studying chain's behaviour at steady state of the network, we suppose that all profiles had been created. Then, if $\mathbb{A}' \notin \mathcal{C}$ at time *t*, in a finite amount of steps the state $\mathbb{A}'$ will be attracted by $\mathcal{C}$. Therefore, on the long term, every states of *S* will be attracted in a finite amount of states by the closed, irreducible and aperiodic class $\mathcal{C}$. As $\mathcal{C}$ is a finite set, all of its states are positive recurrent. Then, restricting the chain $\{\mathcal{A}_t\}$ to $\mathcal{C}$, it is closed, irreducible, aperiodic and every state is positive recurrent, hence it is ergodic.

The ergodic property, ensures limit distribution existence. In this case, if $\mathbb{A} \in \mathcal{C}$,

$$\begin{aligned}\pi_{\mathbb{A}}(\infty) &= \lim_{t\to\infty} P(\mathcal{A}_t = \mathbb{A}) \\ &= \lim_{t\to\infty} P\left[\bigcap_{f\in\mathcal{F}_\infty}\left(\{\alpha_t(f,f)=1\} \cap \bigcap_{\substack{g\prec f \\ g\in\mathcal{F}_\infty}} \{\alpha_t(f,g) = \mathbb{A}_{f,g}\}\right)\right].\end{aligned}$$

Moreover, friendship functions between profiles *f* and *g* of $\mathcal{F}_t$, α_t, were defined as dicotomic variables taking zero or one according to whether they were Facebook friends or not at time *t*. Then, under the ergodic distribution, as time tends to infinity, random variable $\alpha_\infty(f,g)$, with $f \prec g$, has Bernoulli distribution.

Under *CT* hypothesis, each profile of $\mathcal{F}_\infty$ engages friendship with any other profile with the same probability, let's say *p*, so that random variables $\alpha_\infty(f,g)$, with $f \prec g$, $f, g \in \mathcal{F}_\infty$ are identically distributed.

Also, this variables keeps direct relation with "image index" between profile pairs Y_t. Under *CT* this index is reduced to a constant plus a random variable. This variables are independent for all $f, g \in \mathcal{F}_\infty$, with $f \prec g$, so image index between distinct profile pairs are also independent. Consequently, $\alpha_\infty(f,g)$, with $f \prec g$, $f, g \in \mathcal{F}_\infty$ are independent.