Alexandre Vinyals Valdepeñas

Diffusion of Innovations in Health Care: Identifying Opinion Leaders and the Network of Influence Between Physicians

TREBALL DE FI DE MÀSTER

dirigit per la Dr. Jordi Duch, Dr. Robert Rallo, Dr. Xavier Guardiola

Master en Seguretat Informàtica i Sistemes Intel·ligents

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2014

Contents

Chapter 1

Introduction

This thesis aims to identify opinion leaders within a health care organization, using both clustering and data mining techniques, applied over a real world medical prescription database, previously anonymized. The analysis of prescription databases provides key ingredients to infer the underlying social structure of the organization and its possible opinion leaders. The information is inferred analyzing behavioral patterns and establishing causality relationships between physicians.

Organizations are complex social entities, their social complexity can be understood with the study of large and complex social networks. Revealing information can be extracted from the link analysis of nodes. Data mining and clustering techniques allow the extraction of behavioral patterns, furthermore infer causality relations.

Having knowledge about the underlying social structure of an organization its beneficial from a business management perspective, specially if opinion leaders are identified. This knowledge could be used to reduce operational costs and achieve results in a more efficient way (Eg. Changing behaviors without targeting all the physicians). In fact, pharmaceutical companies manage to obtain this information, and use it in their marketing campaigns. Allowing them to specifically target opinion leaders among the physician community. Convincing those opinion leaders to adopt their drugs increases the chances of a faster adoption rate for their drug.

New medical drugs behave as innovations, and follow diffusion processes as well. Accessing the medical prescription database allows to perform detailed analysis of those

diffusion curves. Applying data mining, clustering and user-profiling techniques raises privacy concerns for physicians.

This document shows a brief overview of which information could be obtained from those databases. Finally, the network of influence is inferred, along with the identification of possible opinion leaders. At the very end, a possible solution to endorse physician privacy is briefly discussed.

Chapter 2

Background Information

This section provides a brief background about diffusion, innovations and opinion leaders.

2.1 Diffusion of Innovations

The literal meaning for diffusion is "*to spread out*", this thesis is based in the analysis of adoption curves presented in diffusion processes. The concept of diffusion is used in a wide array of academic subjects, mostly known in physics and chemistry, where diffusion processes describe the motion of substances from high concentration areas to less concentrated areas. But the concept is applied in other disciplines as well, such as biology, sociology, finance and economics.

Everett M. Rogers, the father of *Diffusion of Innovations* [10] describes diffusion as the way in which an innovation is communicated through certain channels over time, among the members of a social system.

Ev. Rogers followed diffusion processes occurring in the agricultural industry. Being born in a rural family he saw how his father loved electromechanical farm innovations, but showed reluctance for biological and chemical innovations. Thus, his father did not adopt a new corn seed which yielded 25% more crop and was resistant to drought, while his neighbor did. During the Iowa drought of 1936, while the hybrid seed corn stood tall on the neighbor farm, the crop on the Rogers farm wilted and his father was

finally convinced [1].

Ev. Rogers categorized the adopters of a new idea or product in the following categories (1) innovators (2) early-adopters (3) early-majority (4) late-majority (5) laggards . His categorization is based in standard deviations from the curve of adopters. Adoption curves are characterized by an *S-Curve*, slow start with an acceleration in intermediate stages, to finally slow-down as it gets leveraged by laggards finally adopting the innovation. Both curves are shown in fig. 2.1, where the light grey curve shows a cumulative percentage of adopters for a given innovation, and the dark curve shows the rate at which new individuals adopt the innovation.

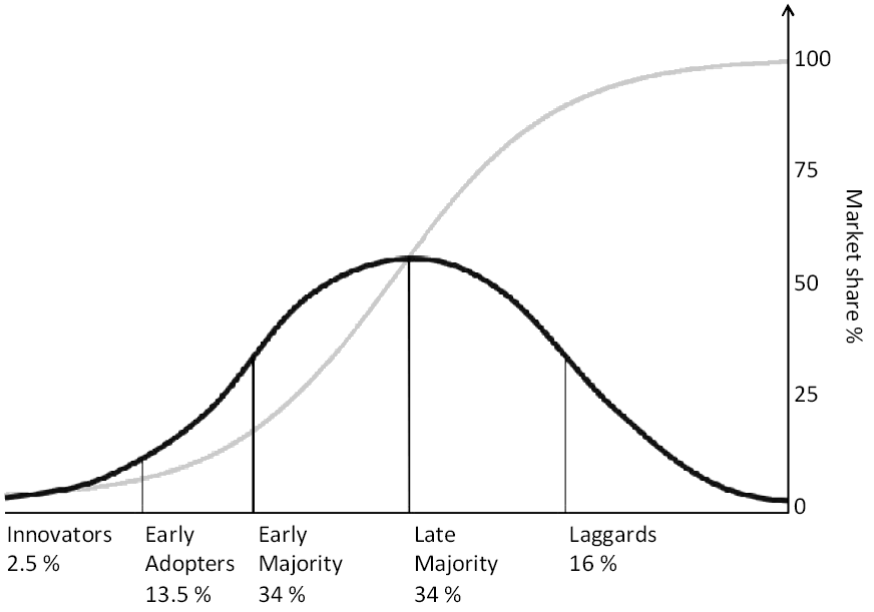

Figure 2.1: Curves of diffusion and adoption

Ev. Rogers defined that key elements in the diffusion process are (1) innovation, as the idea, practice or product that is perceived as new by an adoption unit (2) communication channels, by which messages get to one individual to another (3) social system, as a set of individuals that share a common goal or problem (4) time, as the period needed for the innovation to spread among the members of a social system .

As far as we are concerned, those factors are represented in a health care organization as (1) innovation, as a new medical drug (2) communication channels, as the doctor-to-doctor network, pharmaceutical marketers and medicine journals (3) social

system, as the physicians working inside the health care organization (4) time, as the time required for the new medical drug to reach a late-majority of adoption. .

2.2 The Role of Opinion Leaders

One key factor contributing to the diffusion of innovations is the influence that individuals have among others. Among this influential people there is what its called Opinion Leaders. Those individuals are found to be quite influential during the diffusion of innovations — whether it is in a positive or negative way —. In fact, opinion leaders are used to accelerate diffusion processes [12] as it is more efficient to target individuals that are ready to adopt innovations, than those who may not be.

Opinion leaders play a major role in the diffusion of information, knowledge and innovations among their niche of followers. For this reason, identifying opinion leaders its important from an organizational management perspective, as they can be effectively used to promote behavioral changes inside organizations [13].

The extraction and automation of this process its interesting from a Business Intelligence perspective as well, transforming raw data into meaningful and useful information for business purposes, creating new opportunities and strategies to provide competitive advantages to business.

Chapter 3

State of the Art

This section provides a global overview of already conducted researches and discussed topics which we consider of interest for the main goal of this thesis. Our interest scope ranges from diffusion of innovations in health care to the privacy concerns introduced by the usage of prescription data mining. For this reason, we provide overviews and insights for such topics, and resources where researchers found evidence of

- Diffusion processes taking place in the health care industry in the form of new medical drugs
- Influence and Peering effect between physicians, affecting the diffusion process of new medical drugs
- Privacy concerns associated with prescription data mining and physician profiling

3.1 Diffusion of Innovations in the Health Care Industry

In the health care industry, innovations are presented as new treatments, surgery procedures, or new medical drugs. Successful innovations become widely used among the physician community, describing what its known as a diffusion process.

A recognized study called *Diffusion of Innovations Among Physicians* [4] used a combination of surveys, interviews and prescription record analysis to understand how relationships between physicians affected the diffusion of *gammanym*, a new antibiotic "wonder drug" with lesser side effects that spread out rapidly among the doctors in the medical community. This study established the importance of interpersonal networks as a communication channel for the diffusion process of innovations.

They used surveys and interviews to gather sociometric data about physicians in the organization. This information allowed them to directly construct the relationship network between physicians. To verify the effectiveness of the network they expected pairs of socially connected doctors to show similar behaviors. With a snowballing diffusion process — as it was occurring with gammanym — that is that socially connected physicians adopted gammanym at about the same time.

This approach implies that simultaneity between pairs its a consequence of social connectivity, as they saw how pairs of related physicians adopted the drug at about the same time. They found how the doctor-to-doctor network operated most powerfully during the first five months of the adoption process, and the peak effectiveness of the social links showed up during the first and second month of the adoption process, presenting a sharp decline in its effectiveness after that.

Doctors who still had not introduced the drug by the sixth month after its release seemed to be unresponsive to the social influence. When those doctors finally adopted it, it was because of external influences such as journals, ads or marketers... but not in response to their relationships with other doctors.

Its also stated how physicians who were deeply integrated in the professional com-

munity presented faster adoption rates than those who were not.

3.2 Peer Effect Among Physicians

Empirical evidence for the presence of peer effects among physicians in a health care organization has been reported in *Is There a Physician Peer Effect? Evidence from New Drug Prescriptions* [8]. This study used prescription databases from a health care organization in Taiwan, specialized in patients prescribed for schizophrenia. They state that to address major challenges for empirical studies of the peer effect, their data-set had unique descriptors for patients, physicians and hospitals during a 14-year period. This allowed them to identify hospital-physician-patient pairs, which is a key factor to identify the peer effect between physicians. They found that peer effects seemed to be stronger for stable groups and that the effect intensified for larger groups, or when those groups were of the same age. This seems consistent with the observation of faster adoption rates for integrated professionals in the community.

The study reports that the measured peer effects where on maximums for newly introduced drugs, which reinforces the idea that the diffusion processes in new drugs are suitable to identify influence relationships. The impact of the peer effect in the context of prescription choices made by physicians has been studied as well in *Asymmetric Peer Effects in Physician Prescription Behavior: The Role of Opinion Leaders* [9], where they specifically looked for asymmetric peer effects. An asymmetric peer effect means that the influence between physicians is directed from opinion leaders to their followers, but not vice versa. They successfully quantified this effect, which manifested in the prescription patterns of physicians.

3.3 Opinion Leadership

Ev. Rogers relied in a *two-step flow of communication* model for his theory, which states that most people form their opinions under the influence of opinion leaders [1].

Innovators are the kick-starters of diffusion processes. Then in the early stages of the adoption process, a group of opinion leaders and early adopters begin adopting the innovation. Opinion leaders play a major role during this process, as they select ideas they would like to try and cope with the associated risks of trying new things.

His role its intensified in situations of high uncertainty, in which they build enough trust on their followers to adopt new ideas [3], contributing to the reduction of knowledge that remains unused, plus making organizations progress. New medical drugs are examples of situations with high-uncertainty, as the drug has not been widely used before. This situations of high uncertainty could be related with noticeable increase in the quantification of the peer effect for new drugs.

In *Disseminating Innovations in Health Care* [2] they state things such as (a) find and support innovators (b) invest in early adopters (c) innovators are diamonds in the rough . They provide arguments that reinforce the importance of opinion leaders and their important role within organizations. Which somehow justifies the efforts spent in their identification.

In fact, physicians who likely are opinion leaders get targeted by pharmaceutical companies, which personally send their marketers in an effort to increase the adoption rate of their drugs, hopefully preventing the prescription of the competence drugs. The reasons they target opinion leaders are obvious, not only they become more efficient in their marketing efforts, but if an opinion leader adopts their new medical drug, chances are that physicians who follow his counsel are more likely to adopt the drug as well.

[1] The model may not be fully accurate, specially since the democratization and wide access to information available nowadays, but even with innovations reaching a wider population through mass media, and less word of mouth, persons still seek for advice from their opinion leader.

3.4 Network Inferring

Disposing socio-metric data allows to directly build the peer network. However, the underlying network in a diffusion process usually remains unobserved, as its the case in this thesis. There is no alternative but to infer such network.

The observation and study of behavioral patterns provided key ingredients for inferring relationships in *Inferring friendship network structure* [5], where behavioral patterns from mobile phone data have been used to identify the underlying social network. As its stated that *"data collected from mobile phones have the potential to provide insight into the underlying relational dynamics of organizations, communities and, potentially, societies"*. Could the study of behavioral patterns on a medical prescription database lead to similar results?

In *Inferring networks of diffusion and influence* [6] they state how inferring a network its possible provided that one has a detailed description about the infection of the nodes of such network. That is, a detailed description with the times in which nodes adopted pieces of information or innovations.

3.5 Prescription Data Mining and Privacy Concerns

Prescription data mining is extensively used by pharmaceutical companies. Its said that when drug company representatives visit physicians to market their products, they already know the prescription patterns of the physician they are visiting [11]. The acquisition of this data its done by buying prescription databases from pharmacies. Medical prescriptions are filled with personal information of the patient along with the prescriber information. Its clear that applying user profiling techniques over physicians and patients raises privacy concerns for them [15].

In 2006 several states in the U.S. enacted laws to ban the use of physician prescription databases for commercial uses. Those laws were fought by large medical data collection firms until 2011, when the laws were finally struck down by the supreme court, alleging a commercial free-speech rights in violation of the First Amendment of the U.S. Constitution [14]. Therefore, physicians privacy its still a concern nowadays. The topic has been discussed in law, medicine and ethics journals, where they suggested how those issues could be alleviated by requiring explicit consent from physicians to receiving salesman visits. It seems thought that the safety of consumers keeps being handed to an unregulated private market [7].

Chapter 4

Data sources

To achieve our goals we used a prescription database obtained from the data warehouse of a health care organization. The database contains prescription records from a population of about 200.000 citizens during a 22 months period, from 2012 to 2013.

First step was to construct a new data-schema, applying extraction-transform-load techniques to the original records, the schema has the following fields (1) Anonymized patient code (2) Anonymized physician code (3) Product code (4) ATC [1] code (5) Prescription date (6) Prescription center code (7) Patient birth yr. .

The medical drugs physicians prescribe are represented in our database with a numeric identifier. The category of this product its stored in the ATC field, which contains a classification nomenclature for medical drugs. When looking for the introduction of new products, we look at product codes appearing for the first time in some time t. A small sample of those records is provided in figure fig. 4.1, along with a numerical description of the entire data-set in fig. 4.2.

[1] The ATC code has 7 characters that can be clustered in different levels *i.e the first character of the code represents the 1st level, corresponding to the anatomical main group of the drug. Three characters represent the 2nd level, corresponding to the therapeutic subgroup of the drug.*

	Patient	Physician	Product	ATC	Price	Date	Center	Birth yr.
1	ee8070d	a9c1d0a	876466	S01XA20	5,39 €	2012/01/01	01522	1931
2	66b92b4	1903b7a	831552	J01CA04	3,09 €	2012/01/01	01327	2000
3	7d52679	c55d1c7	837047	N02CA51	1,5 €	2012/01/01	00705	1950
4	43a09ca	49e5d85	815241	M01AX25	19,37 €	2012/01/01	00705	1940
5	a6525c3	f4ad493	906214	C05CA04	10,79 €	2012/01/01	00705	1952
6	742b032	a732122	788927	R05CB01	2,15 €	2012/01/01	00705	1921
				...				

Figure 4.1: Sample with prescription records of the database

		$\bar{x}$	σ	min	max
N. of prescriptions	5,192,620	-	-	-	-
N. of valid prescriptions[2]	4,924,855	-	-	-	-
N. of unique products	9,310	-	-	-	-
N. of unique ATCs	966	-	-	-	-
N. of months	22	-	-	-	-
Unique patients	186,179	-	-	-	-
Unique physicians	508	-	-	-	-
Patients for physician	-	971	1,926	1	29,303
Products for ATC	-	10	21	1	196

Figure 4.2: *Database descriptive attributes*

The high deviation of *patients treated by physician* suggest significant differences in the volume of patients that each physician treats. Extracting the distribution of unique patients treated by physician in fig. 4.3 confirms those differences. The distribution confirms the existence of a consolidated and integrated professional community, which handled in 20 months more than 1,000 unique patients.

[2]Valid prescriptions stands for prescriptions with all the required fields, *ie. prescriptions without a product code are not accepted*

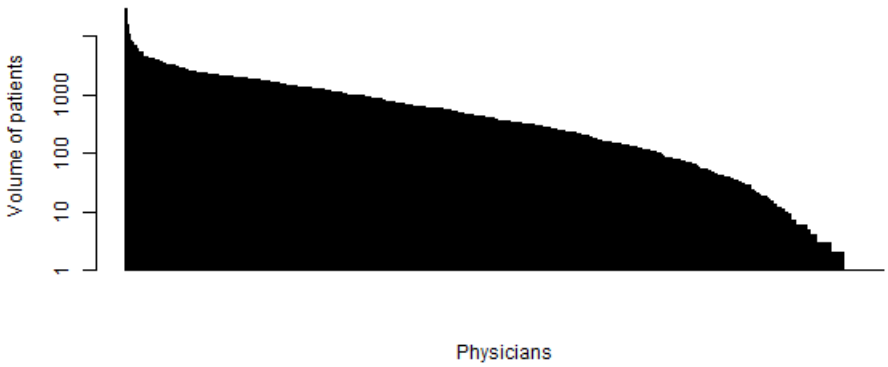

Figure 4.3: Distribution of patients treated by physicians

The same histogram is extracted to see how much products are classified in each ATC classification code, the distribution in fig. 4.4 shows significant differences again. Large amount of products remains classified in a single ATC. This could be explained by the drug being presented in different ways (Eg. different number of pills or different concentrations), but the reason could be a high competence in the market as well. The ATC categories with more products are for gastro-esophageal re-flux diseases, and antibacterials derived from penicillin.

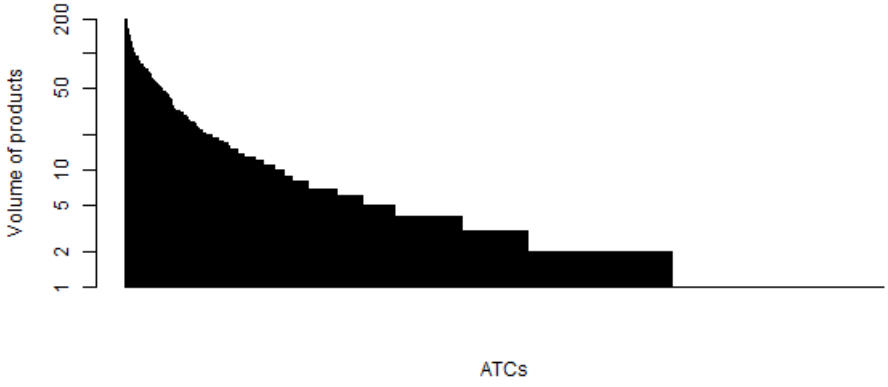

Figure 4.4: Distribution of products in ATC categories

4.1 Data correctness

This section provides a better understanding of the data-set, along with hints on the validity of the data. Also demonstrates which kind of information one could generally obtain mining a prescription database.

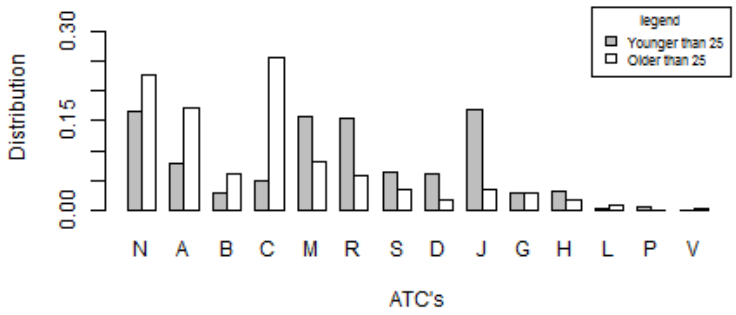

Figure 4.5: ATC distribution comparison for both groups of ages

We extracted the ATC distributions for patients, aggregating them in two major groups (a) patients under 25 years old (b) patients over 25 years old . Different distributions are expected for both groups, as patients in different ages are treated for different issues. A direct comparison of both distributions its shown in fig. 4.5.

For patients over 25 years old, (1) C - cardiovascular (2) N - nervous system (3) A - alimentary tract and metabolism are the most prescribed categories. Whereas patients under 25 years old showed more prescriptions for categories (1) M - musculo-skeletal system (2) R - respiratory system (3) J - antiinfectives for systemic use.

Cardiovascular related prescriptions represent only a 5% of prescriptions for people under 25, quite the contrary for older people, reaching a 25% in the distribution. The same phenomena is observed for the J ATC, the anti-systemic drugs category — which includes vaccines —. The comparison shows how the younger group presents a 16% proportion for J, whereas old people hardly reach a 5%. This seems consistent and valid, as one would expect expect people under 25 to receive more vaccines in proportion to other drugs, than the group of people over 25 years old.

In a similar way that patients in different age groups get different prescriptions, the professional specialty its expected to be reflected as well in the prescribing patterns of physicians. First hand information from the organization told us to expect two major groups of physicians (a) 70% of family physicians (b) 30% of pediatricians .

To visualize those differences we extracted the prescription patterns for all physicians, showing the distribution of ATCs they prescribe. The patterns are visualized as a heatmap in fig. 4.6, rows representing different physicians and columns representing different categories of drugs, clustered by the first level of the ATC.

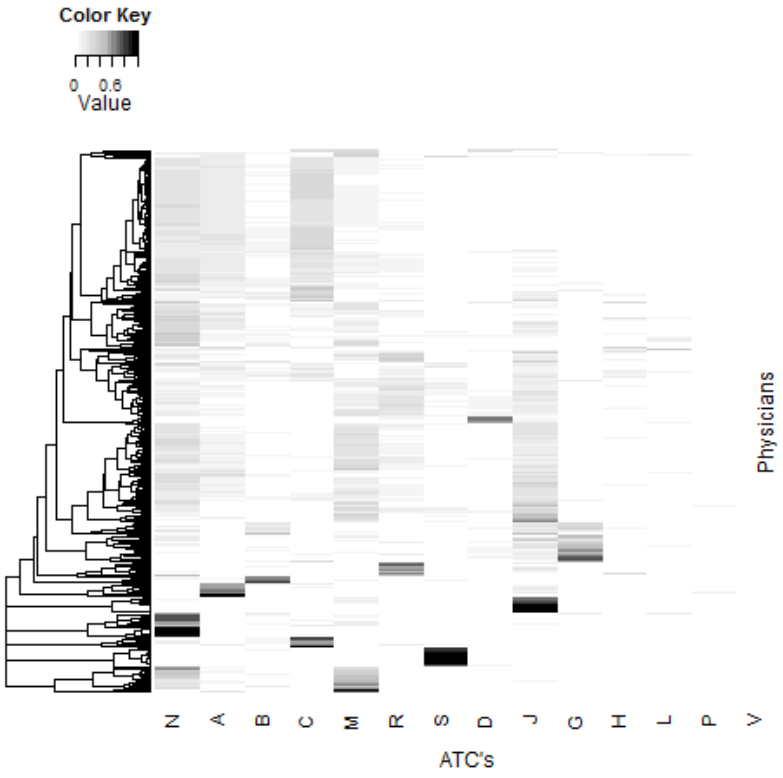

Figure 4.6: Prescription patterns for physicians

The differences in prescription patterns are clear. There exist small groups of physicians that prescribe almost entirely in one or a couple of ATC categories (Eg. physicians prescribing only in the S category, corresponding with sensory organs such as eyes and ears), they are clear examples of physician specialty at its best. The small group of physicians mostly prescribing drugs belonging to the S category probably are ophthalmologist or otologists.

Cutting the physicians tree by the fifth level resulted in one of the clusters containing 360 physicians, a size representing about 70.81% of the physicians in the organization, a value we were already expecting.

Chapter 5

Analysis

At this point the dataset set has been described and characterized, we looked at how products, patients and ATCs are distributed along with the prescribing pattern differences between physicians.

Previous literature about diffusion in health care and peer effects between physicians reassured us that the effects are measurable and quantifiable. But perhaps more importantly, that they have been obtained analyzing prescription databases.

This thesis infers influence relationships between physicians analyzing the diffusion curves presented in the adoption of new medical drugs. Related work stated how peer effects accelerate the diffusion process, and how the doctor-to-doctor effects are noticeable until the fifth month. Also how socially related physicians are expected to introduce the drug at about the same time, suggesting simultaneity as a relevant factor to identify the social relationships.

Our prescription database contains a detailed description for any diffusion process unfolding over time[1] during a 20 months period. This allowed us to infer causality relationships between physicians, inferring the cascading of influence trough the physicians network. To do this, the diffusion curves for new medical drugs must be previously identified from the raw data.

[1] Within one month resolution

Related literature focused in one or a few new medical drugs, with heavy support from socio-metric data that allowed them to directly observe the underlying social network of the diffusion process. Given that we do not dispose of socio-metric data, the underlying network for this study remains unobserved. In an attempt to improve the reliability of the results, we performed an aggregation of inferred results from individual curves.

With the aggregation process, its expected that early adopters who happen to be opinion leaders, consistently adopting new drugs, will eventually get differentiated, as they will present more inferred relationships than physicians who are not, successfully identifying physicians who are likely to be opinion leaders, within the organization.

5.1 Identifying Diffusion Curves

Before proceeding, we must figure a way to identify new products and their adoption curves inside the raw dataset.

The straight forward way to identify diffusion curves is to look at the evolution of how many physicians prescribe a drug over time. Given that the data-set has prescription records from a 20 months period, we must ensure that seasonality does not produce false positives (Eg. products appearing in winter due to colds would describe an adoption curve, yet they would not be an innovation or new drug), for this reason the identification of new drugs will be performed for products appearing after the first twelve months, to avoid one-year seasonality effects. This simple restriction reduced the search scope from 9,310 to 1,103 products.

We must ensure more restrictions to hold true in order to identify diffusion curves. The problem its that those restrictions are based in *physicians over time* and the direct extraction of product histograms leaves us with *volume of prescriptions over time*. The volume of prescriptions does not represent meaningful information as far as a diffusion process its concerned, a transformation from *prescription volume* to *physician volume* has to be done. This transformation its done by extracting subsets of physicians under the curve at each temporal step. After transforming the histograms for each product,

now we can apply restrictions considering prescriber volume.

A newly introduced product will follow the conditions shown in eq. (5.1). Where $A(p,t)$ stands for the number of adopters prescribing a product p in time t, finally m_i stands for the introduction month of the new drug, that is the month in which the number of adopters prescribing p its greater than 0 for the first time in the entire period.

$$\begin{cases} A(p,t) = 0, & \text{for } t \in [0, m_{i-1}] \\ A(p,t) > 0, & \text{otherwise} \end{cases} \tag{5.1}$$

When extracting adoption curves we would like to control (1) granted length of the curve (2) minimum number of adopters .

1. To grant a minimum length of n months, we check that $A(p,t) > 0$ holds true during the n previous months before the last available month in the data-set, the twenty-second month M_{22}.

2. A minimum of k adopters its granted restricting $A(p,t) > k$ for $t = M_{22}$

Tuning different combinations of parameters, a set of scenarios its defined in fig. 5.1, each one accompanied with the amount of curves identified for such scenario.

	Min. len	Cond. min. len	Min. adopters	N. curves
Scenario 1	2 mos.	$A(p, M_{20}) > 0$	$A(p, M_{22}) > 20$	58
Scenario 2	2 mos.	$A(p, M_{20}) > 0$	$A(p, M_{22}) > 10$	122
Scenario 3	2 mos.	$A(p, M_{20}) > 0$	$A(p, M_{22}) > 0$	393
Scenario 4	5 mos.	$A(p, M_{17}) > 0$	$A(p, M_{22}) > 20$	23
Scenario 5	5 mos.	$A(p, M_{17}) > 0$	$A(p, M_{22}) > 10$	42
Scenario 6	5 mos.	$A(p, M_{17}) > 0$	$A(p, M_{22}) > 0$	186

Figure 5.1: Applied restrictions and results

When we previously analyzed the prescription patterns of physicians in fig. 4.6, the smaller clusters had in average a size of 20 physicians. This makes $k > 20$ a reasonable restriction in terms of adoption widespread. Higher values would exclude specialized physicians, as a new product could not reach more physicians than themselves.

The next logical restriction is no restriction at all with $k > 0$, and for completeness we add an intermediate step with $k > 10$.

For the temporal length of the window, the reasonable choices are two and five months: two months being the number of months in which the doctor-to-doctor network effectiveness its at his peak and five months as it is the period in which the doctor-to-doctor network effects are still noticeable. Examples of the identified curves are shown in fig. 5.2.

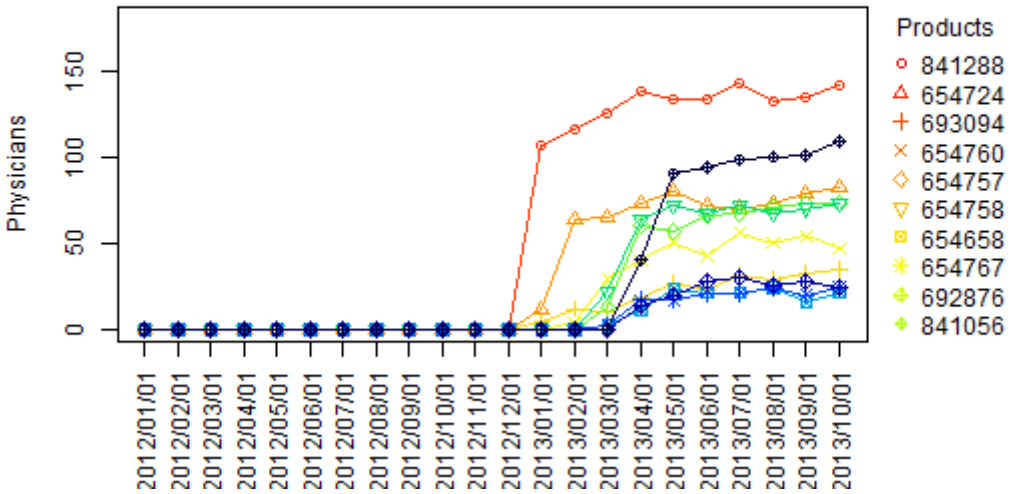

Figure 5.2: Sample adoption curves from the database

At this point one can extract the subset of physicians under the curve at any given point in time — within the one month resolution — which allows us to proceed to the next step of inferring influence relationships between physicians.

5.2 Inferring Relationships

After identifying the diffusion curves we can produce a model to store the data of the curve in a more usable way, like in fig. 5.3. The idea is to build a table representing the presence of the physician in each month of the curve. Rows represent the physicians under the curve, while columns the prescription volume of such physician at each month[2] of the window. Physicians who join the curve past the first month are preceded by zeroes, but for reading clarity they are marked as - to clearly show when each physician joins the curve. This table is constructed for all the available curves of each scenario.

	Physician	M1	M2	M3	M4	M5
1	ee3c678ee95a56a7b942dd00078b76e3	1	0	0	1	0
2	b16763ffe47918c479bc1104f0aab00d	22	43	99	91	46
3	a99e1fa6e1eee294f05d6ebf255f20b3	1	1	3	3	2
4	2dd7840f5474d7cd648c8dfb4db68a8b	-	8	2	10	2
5	278c3201e9ffacb3c1f30eac62a420d0	-	-	1	1	0
6	aaa74bef4ea7f6b7def4bc7538da45e3	-	-	1	0	0
7	f3147f883b433fd2d39c6a273e6fe911	-	-	-	1	0
8	2ddc2e49b5ebd6da78001be1560a741a	-	-	-	6	22
	...					

Figure 5.3: Sample of a histogram table for one specific product using a five month window

Each temporal step or column has a subset of adopters who join the curve. In the invented sample from fig. 5.3 physician 4 is a new adopter for the second month. Physicians 1, 2 and 3 are adopters in the first month. If we consider for this experiment that the only communication channel available for the diffusion of innovations its the doctor-to-doctor network, the only way for physician 4 to join the curve would have been to get influenced by one — or more — of the physicians 1, 2 and 3. Basing the process in causality, physicians 1, 2 and 3 would launch arrows to physician 4 indicating an inferred connection between them. In second month, arrows would be launched from physicians 2, 3 and 4 to physicians 5 and 6 — physician 1 does not launch arrow as he is no longer present in the process by the second month —. The last arrows would be

[2]We used prescription volume to make the model more descriptive, ideally its enough with a boolean representing whether the physician prescribed or not, for each month.

launched from the physicians under the fourth month to the new adopters in the fifth month.

This results in the construction of what we call *connectivity* matrix, along with a simultaneity matrix. Examples of both matrix are shown in figs. 5.4 and 5.5 for the sample table in fig. 5.3, where columns and rows represent physicians.

	1	2	3	4	5	6	7	8
1	0	1	1					
2	1	0	1					
3	1	1	0					
4				0				
5					0	1		
6					1	0		
7							0	1
8							1	0

Figure 5.4: Sample simultaneity

	1	2	3	4	5	6	7	8
1	0			1				
2		0		1	1	1	1	1
3			0	1	1	1	1	1
4				0	1	1	1	1
5					0		1	1
6						0	1	1
7							0	
8								0

Figure 5.5: Sample connectivity

The connectivity matrix represents whether one row physician p_i could had any influence over the column physician p_j. The process may be dull for a single adoption curve, however as its been shown in fig. 5.1 we dispose of a great number of adoption curves. The aggregation of those matrix for each curve will eventually differentiate those physicians who differ from the rest.

Chapter 6

Results

This chapter presents the results obtained from the inferring process. The results are presented as heatmaps, clustered by euclidean distance. The clustering hierarchy obtained from the simultaneity matrix its forced on the connectivity heatmap, allowing for a direct comparison of both heatmaps, as dendrograms retain the same order. Both the columns and rows of the heatmaps represent physicians, while the value of each cell represents the weight of the relationship, which has been obtained from the aggregation of all the available curves.

6.1 Clustering methods

Figures 6.1 and 6.2 show the clusters formed in the simultaneity matrix, using a 5 months window with 23 curves and at least 20 adopters, in both euclidean and cosine dissimilarity distances.

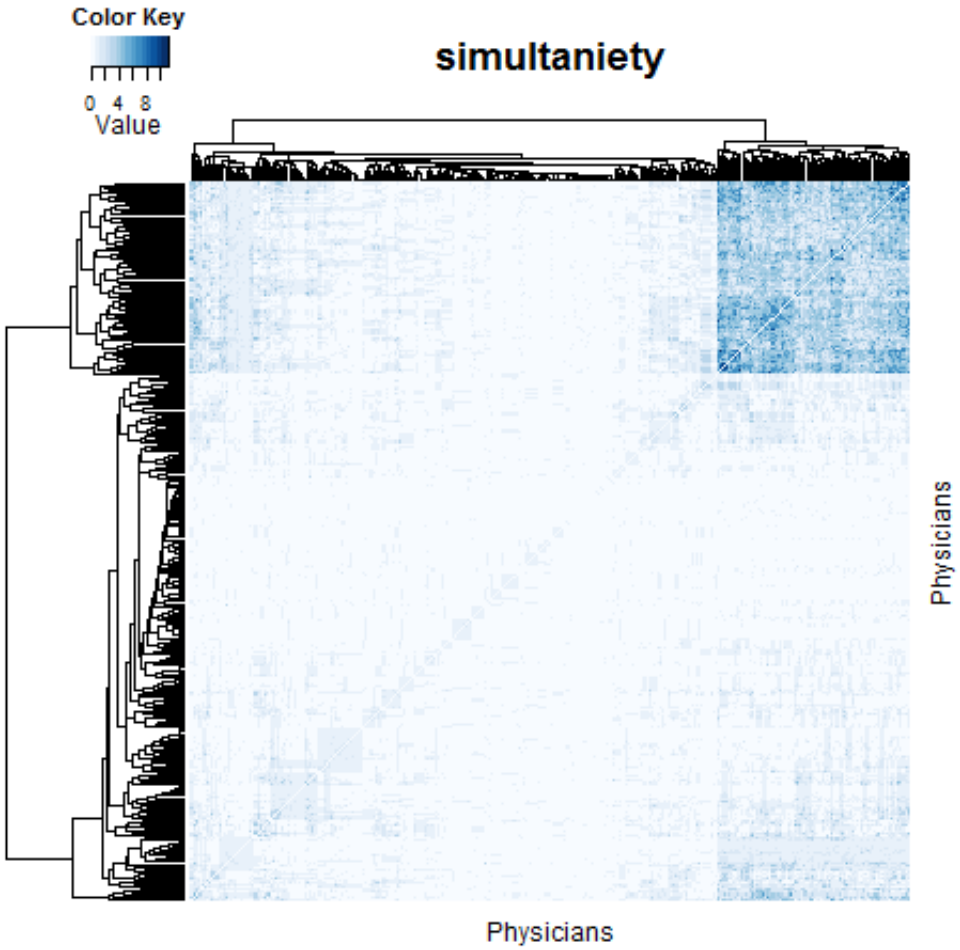

Figure 6.1: Euclidean distance

When extracting the clusters, it becomes handier to use euclidean distances, as possible opinion leaders get clustered at the highest levels of the hierarchy, whereas in cosine dissimilarity the groups are in way deeper levels, making it more tedious to handle. Also, euclidean distance provides a crisp differentiation of the clusters, which seem to fade when using cosine dissimilarity. The upcoming section will show the results using only euclidean distances.

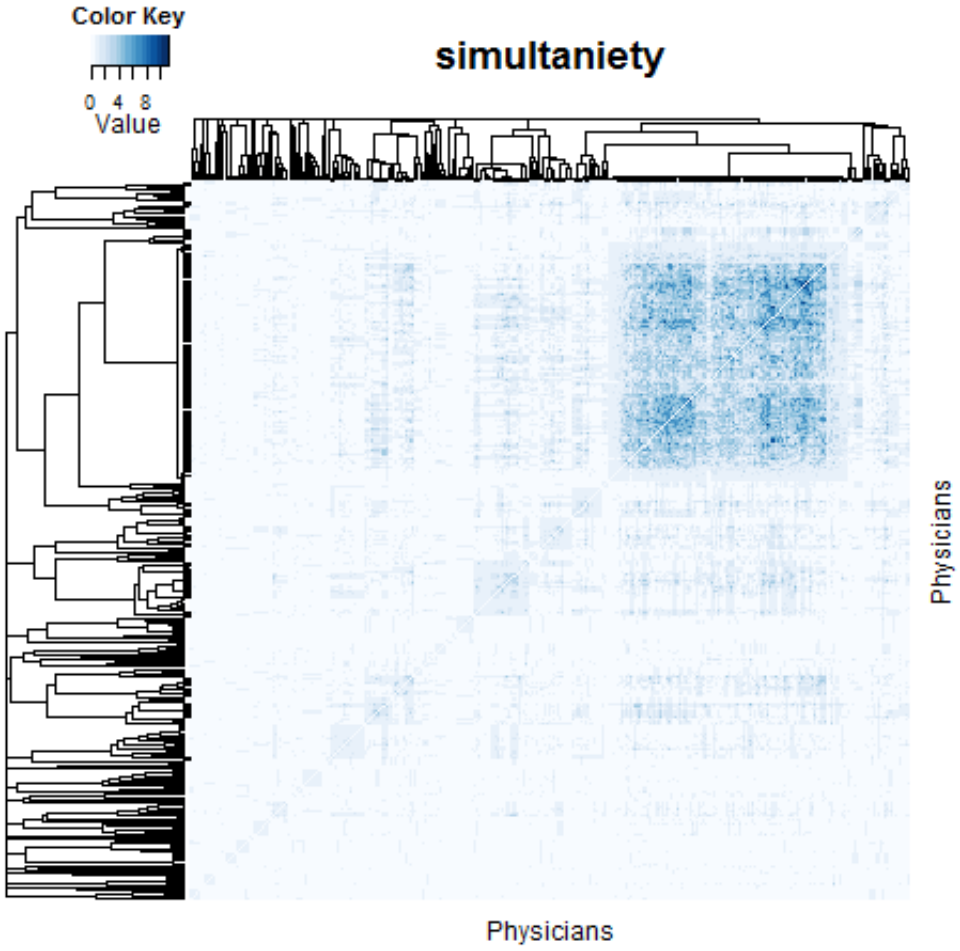

Figure 6.2: Cosine dissimilarity

6.2 Scenario comparison

6.2.1 Window length

This section shows how results vary depending on the window size. Figures 6.3 and 6.4 show simultaneity heatmaps for scenario 1 and 4 (2 and 5 months window). Figures 6.5 and 6.6 shows the respective inferred connectivity.

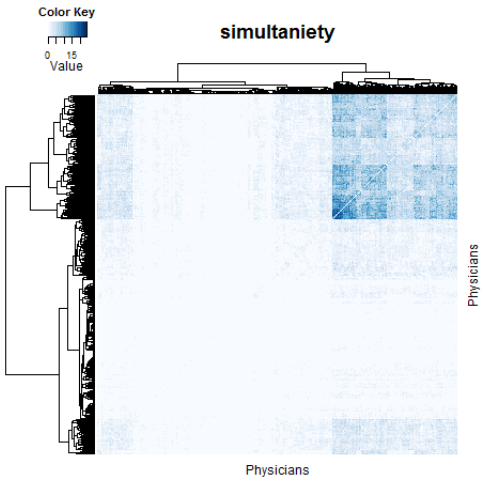

Figure 6.3: Simultaneity using 2 months window

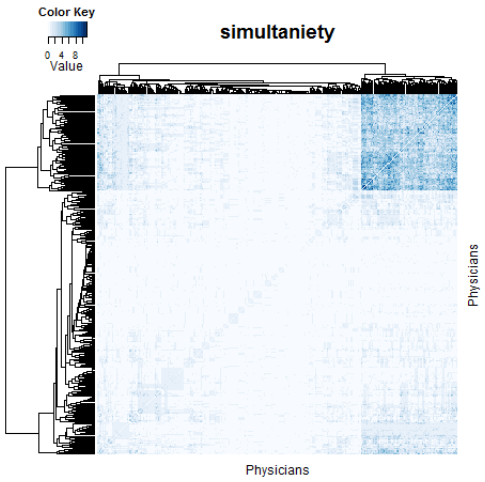

Figure 6.4: Simultaneity using a 5 months window

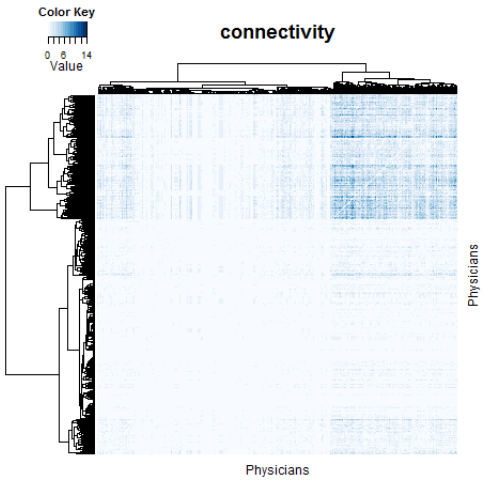

Figure 6.5: Connectivity using 2 months window

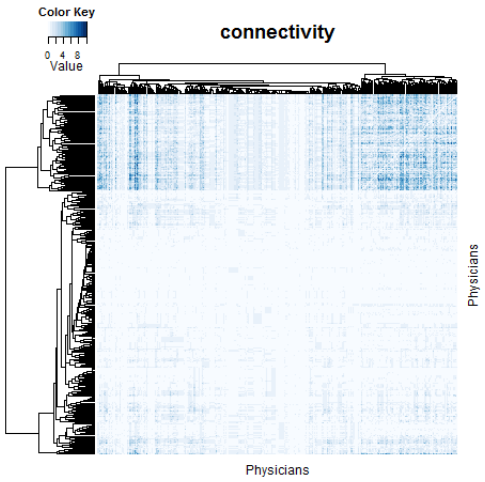

Figure 6.6: Connectivity using a 5 months window

The effects of using a wider window are notorious, inferring in a 5 months window has extra depth of inferring, which is noticeable as the connectivity reinforces horizontally, showing relations that were not inferred in the 2 months window. However, using a 2 months window allowed to better differentiate several clusters in the top-right area of the map in fig. 6.3. The reason behind this behavior is that using a 2 months window not only yields more curves (more aggregation, better inferring) but uses data in the most effective period for peering effects, obtaining more precision and less noise, unlike in the 5 month depth inferring.

6.2.2 Volume of adopters

We have seen how the length of the window affects the results. Now we present the differences that occur when curves with less than 20 adopters are used during the process. Figures 6.9 and 6.10 show the differences for a 2 months window and both $k > 20$, $k > 0$. Their respective simultaneity its shown below in figs. 6.7 and 6.8.

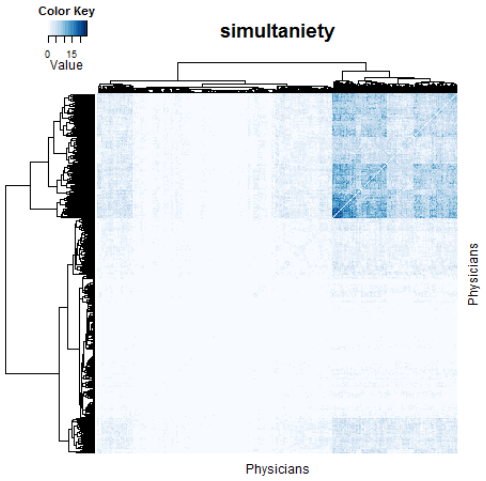

Figure 6.7: Simultaneity with > 20 adopters

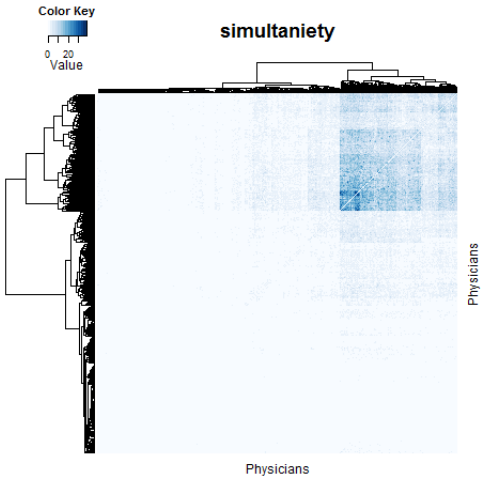

Figure 6.8: Simultaneity with > 0 adopters

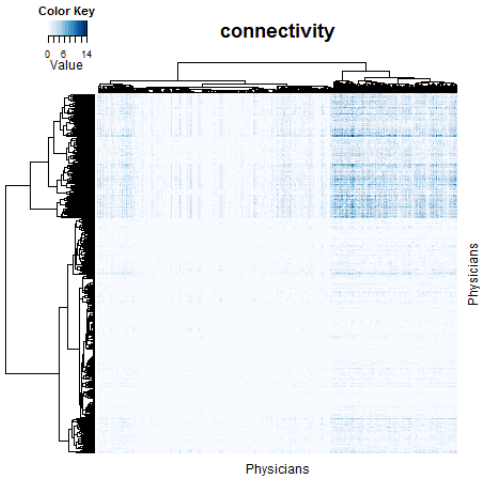

Figure 6.9: Connectivity with > 20 adopters

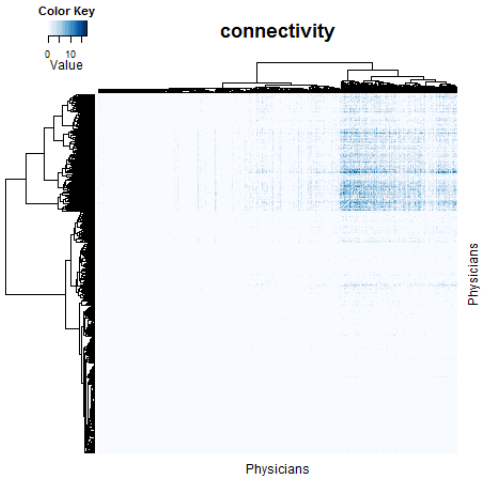

Figure 6.10: Connectivity with > 0 adopters

The implication of using $k > 0$ is that curves with small amounts of participants contribute to the aggregation process. One could think that small curves with small amounts of physicians involved would not produce significant differences, yet if those small curves involve opinion leaders, the result is what we see in fig. 6.8, that those

who are likely to be opinion leaders get differentiated from the rest. Looking at the range of values, using $k > 0$ augmented the maximum value in the scale from 25 to 40, which means that those small curves involved physicians who already had significant values of simultaneity. Opinion leaders are expected to try new things, hereby the small curves as they involve mostly opinion leaders themselves. The $k > 20$ scenarios use curves in which the diffusion and widespread of the drug considerably succeeds, which somehow pollutes the inferring process as more physicians who act as followers adopt those drugs. However, removing this restrictions does an excellent job in differentiating possible opinion leaders.

The same is seen for connectivity heatmaps, where $k > 0$ increased the scale from 14 to 20, meaning that already strong relationships got reinforced in value.

6.2.3 Simultaneity and connectivity correlation

The main differences between connectivity and simultaneity are explained by (a) simultaneity is symmetric whereas connectivity looses the symmetry due to the implicit direction of the influence between physicians (b) simultaneous groups of physicians not necessarily influence others, as there could be a simultaneous group of late-adopters joining the curve. However, we can still measure the correlation between connectivity and simultaneity for the different scenarios. In figs. 6.11 and 6.12 both Pearson's and Spearman's correlation coefficients are presented for each scenario.

Scenario			Correlation summary			
mos.	adopters	curves	$\bar{x}$	σ	min	max
2	> 20	58	0.31	0.25	-0.19	0.77
2	> 10	122	0.34	0.26	-0.18	0.80
2	> 0	393	0.36	0.27	-0.18	0.83
5	> 20	23	0.04	0.18	-0.34	0.51
5	> 10	42	0.08	0.20	-0.32	0.60
5	> 0	186	0.12	0.22	-0.30	0.70

Figure 6.11: Pearson's correlation between connectivity and simultaneity

Previous literature told us that simultaneity was noticeable between pairs of physicians during the first and second months of the diffusion process, periods in which

Scenario			Correlation summary			
mos.	adopters	curves	$\bar{x}$	σ	min	max
2	> 20	58	0.33	0.25	-0.20	0.75
2	> 10	122	0.35	0.26	-0.19	0.76
2	> 0	393	0.37	0.26	-0.18	0.79
5	> 20	23	0.06	0.20	-0.34	0.56
5	> 10	42	0.10	0.21	-0.32	0.60
5	> 0	186	0.13	0.23	-0.30	0.62

Figure 6.12: Spearman's correlation between connectivity and simultaneity

connected peers adopted the innovation at about the same time. Also, during the next months the effects of the doctor-to-doctor network were noticeable but not at its peak effectiveness, which weakened the simultaneity between peers.

Similar results are manifested looking at the correlations. In the two months window there is an overall moderate relationship, with slightly strong relations appearing in maximum cases. However, with five months windows the relationships are shown to be weak overall, presenting slightly moderate relationships in maximum cases. The overall conclusion seems that 2 months windows perform a better job to identify opinion leaders, combined with no restrictions in widespread of the drug.

6.3 Differentiating Opinion Leaders and Network Inferring

In the previous heatmaps there was one big cluster of physicians which clearly differed from the rest. This section focuses on that cluster, as it is expected to contain possible opinion leaders. As we have seen, a 2 months window with no restrictions seems to favor the differentiation of physicians, so this section focuses in scenario 3 (with its 393 curves). Applying a connectivity threshold filters about 2/3 of the total physicians. Given that the majority of physicians had none or low values for connectivity, the heatmaps become way smaller.

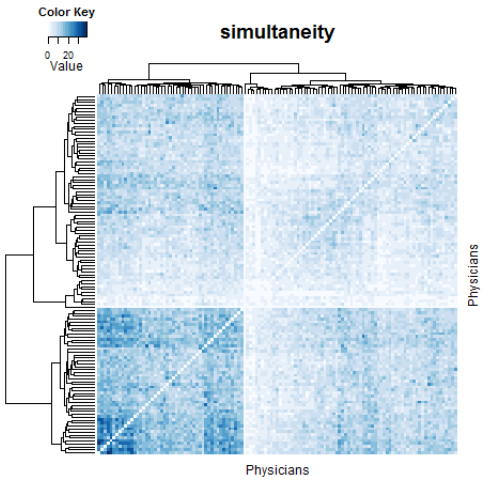

Figure 6.13: Simultaneity with > 0 adopters

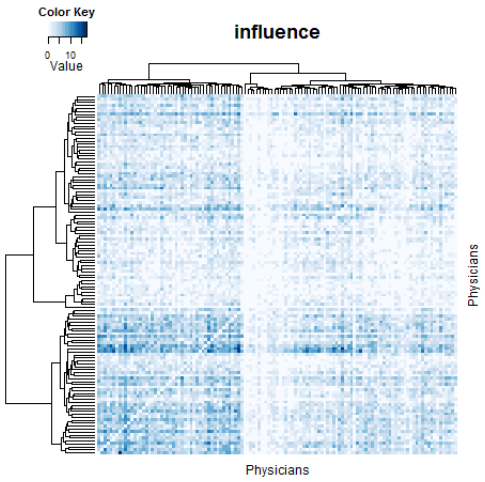

Figure 6.14: Connectivity with > 0 adopters

In fig. 6.13 the big cluster of physicians in bottom-left area shows how indeed simultaneity does not imply connectivity, as the same area in fig. 6.14 presents weak relationships where simultaneity was strong. Also in the mid-top area of connectivity, where simultaneity was not *that* strong, appeared strong lines indicating very influential physicians.

Connectivity reached values around 20 in maximum cases, a reasonable threshold to obtain the possible opinion leaders could be to filter out physicians with a connectivity value smaller than 12. The filtered result is shown in fig. 6.15.

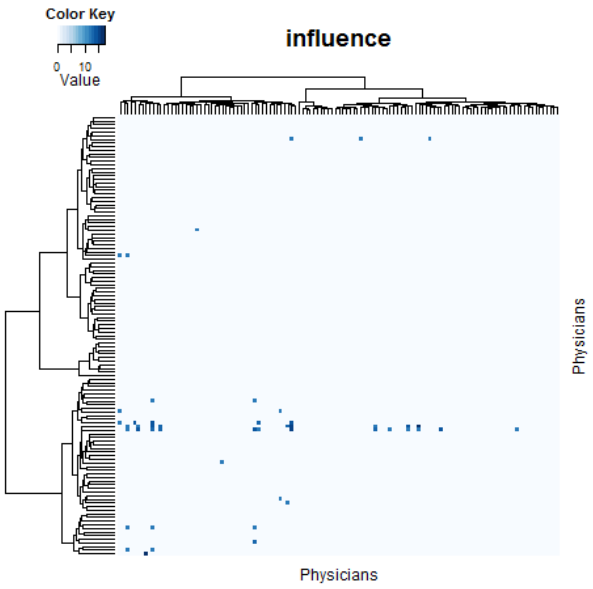

Figure 6.15: Filtered heatmap, connectivity > 12

This adjacency matrix already provides structural information about the network. The inferred network its shown in fig. 6.16. For this network to become usable, the only step left would be to unveil the real identity of the physicians behind its hashed ID.

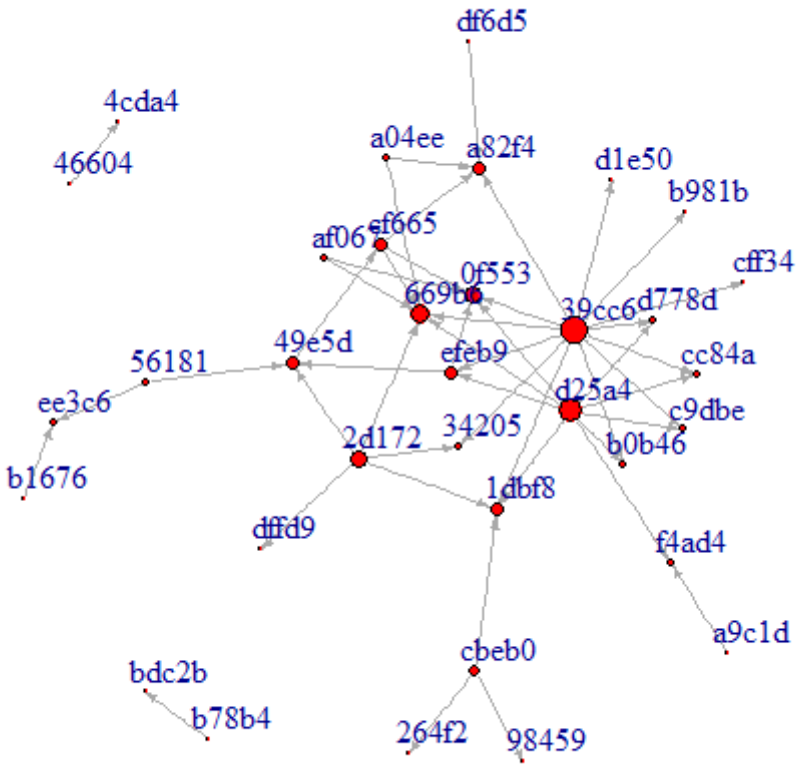

Figure 6.16: Inferred network of possible opinion leaders

Figure 6.16 does more than identifying opinion leaders, it provides information about the existing connections between them. For instance, the node *"399cc6"* with 11 links — all weighted above 12 —, seems to be quite a relevant and socially connected physician inside the organization, specially since all of his arrows point outside of the node, which means that his influence overcomes the received influence from his connected nodes.

The high threshold produced two isolated groups, what could be an indication of physician specialty, or resulting due to physicians working in isolated or different centers, which difficulties the integration with other professionals of the company.

Chapter 7

Protecting Physician Privacy

The profiling of physicians raises privacy concerns. A physician working in a health care organization probably knows and accepts that medical related data could be used by the organization, but remains unclear whether he is fine with his prescribing patterns being handed to pharmaceutical companies, without his explicit consent. This is a direct consequence of pharmacies being able to build prescription databases from dispensed prescriptions.

Anonymizing the physician field in the prescriptions its not a valid solution, as prescriptions are required by law to be prescriber-identified. A system involving government ruled infrastructures could provide a solid foundation along with electronic prescriptions, to enforce the protection and privacy of such information, while increasing the quality of health care due to the advantages of e-prescriptions.

In fact, having an electronic prescription infrastructure has been designated as an important strategic policy to improve health care in Europe, with projects like *epSOS*. However due to the myriad of challenges presented in a system like that, granting the inter-operability between countries, makes it hard to see this system being fully inter-operable and deployed in the near-future.

Private companies such as Surescripts or eRx offer e-prescribing services to improve the quality of health care, yet the issue remains the same, as sensible data is being handled by private corporations. While helpful, one could do better in order to protect privacy. The scope of this paper focuses in physicians, thus we propose a solution

based in smart-cards and signatures to improve their privacy. First with an individual signature schema and finally a better solution involving a group signature schema. Both solutions require of

1. Trusted central authority (CA) to handle key distribution and revocation.

2. Smart-cards for physicians and citizens distributed by the CA.

7.1 Individual signature schema

Given a group of physicians $P = \{p_1, ..., p_n\}$ the CA distributes $K = \{kp_1, ..., kp_n\}$ key pairs such that each $kp_i = (k_i, pk_i)$ is unique in K. The private key k is securely stored in the smart-card, which is finally given to his legitimate owner. Citizens could obtain his smart-cards in a similar way, but they would only require to store his identity number.

A physician and a citizen using this system would both introduce their smart-cards in a device such that the prescription of an ATC α to a citizen C would remain stored in C smart-card with at least the following fields,

- α ATC.
- pk public key of the physician.
- Signature with the private key of the physician $S = E_k(\alpha, C_{id})$

The pharmacy dispenser would request to the CA whether pk is a valid key. In such case, the signed content is recovered decrypting S with pk. The dispenser verifies that the signed fields α, C_{id} match with both the prescribed ATC and the identity of the citizen stored in the smart-card.

With this schema, the prescriber remains anonymous as long as the identity behind the public key pk is not disclosed. An eventual disclosure of the real identity behind pk discloses previous records associated with this key, due to linkability issues. While functional, this schema presents weaknesses.

7.2 Group signature schema

A better system involving a group-based signature schema would solve the weakness of individual signatures, providing real anonymity while keeping prescriptions verifiable by dispensers. In this schema physicians could individually sign prescriptions using his private and group public key. A dispenser could verify the validity of the prescription checking that the signed content was generated by any member of the group. Given that the requirement for belonging to this group is being a physician, the prescription would be valid. And in case of fraudulent prescriptions, the identity of the group member that signed the prescription could be disclosed by the group manager.

Thereby, a set of physicians $P = \{p_1, ..., p_i\}$ is distributed in several groups $G = \{g_1, ..., g_i\}$, such that each group has $M = \{m_1, ..., m_j\}$ members. The group has a public key Y_i, several private keys X_{ij} for each member and the X_{ir} key for the group revocation manager.

The physician smart-card would contain both his private key X_{ij} and the group public key Y_i. A physician and a citizen using this system would both introduce their smart-cards in a device such that the prescription of an ATC α to a citizen C would remain stored in C smart-card with at least the following fields,

- α ATC.
- Y_i public key of the group.
- Signature $S = E_{(X_{ij},Y_i)}(\alpha, C_{id})$

The pharmacy dispenser would request to the CA whether Y_i is a valid group key. In such case, a verification algorithm accepting (α, C_{id}), S and Y_i would return 1 only if signature S was generated by any group member using (α, C_{id}), X_{ij} and Y_i. If a situation required to reveal the real identity of the prescriber, an algorithm using as input the signature S, the group revocation key X_{ir} and group public key Y_i would return the identifier of the group member m_i who issued the signature.

In this schema physician identity remains protected behind the group signature, specially if the groups are large enough to provide enough uncertainty. The linka-

bility issues dissuade as private keys of physicians are not announced, while keeping prescriptions verifiable and the ability to reveal the prescriber identity if needed.

Chapter 8

Conclusions

This thesis showed the entire process followed to infer the influence network between physicians of a health care organization. Furthermore, the identification of opinion leaders among those physicians. The inferred network showed the direction of the influence between individuals. All the results were anonymous as identities were previously masked with a secure hashing function. Though the organization could unmask identities to make this information usable for a wide range of applications in business intelligence (Eg. one could promote or assign team leaders).

Its been shown how the information provided in adoption curves for new medical drugs, allowed us to infer influence relationships between physicians. As the influence matrix for several new drugs adds up, those physicians who act as opinion leaders eventually become differentiated from the rest, making the aggregation process a key factor in the retrieval of meaningful results.

However, external organizations such as pharmaceutical companies could get in hold of prescription databases, and perform similar procedures with non anonymous data. Such thing would compromise the privacy of physicians and patients. Previous literature showed that those privacy concerns are real and remain unsolved as of today, given that pharmaceutical companies acquire prescription databases from pharmacies.

They would face additional problems during the inferring process, as we already knew that physicians belonged to the same community — the health care organization —, pharmaceutical companies would require an extra step to obtain the physicians

workplace — which surely do, as they get to they visit them on their offices —. Applying those techniques would provide a set of opinion leaders to whom market new drugs. Additionally, they could obtain the prescribing patterns and plenty of details about a physicians prescribing behavior.

For this reason, the proposal of an electronic prescription system, supporting group-based signatures, would effectively prevent pharmacies from building prescriber-identified databases, as the identity of physicians would remain anonymous behind the group public key, preserving the legitimacy validation for prescriptions. The identity of the signing member could be revealed in cases requiring it, with intervention of the CA or the designated group leader, which holds a secret key to unveil member identities behind group signatures.

Chapter 9

Acknowledgments

This thesis could have not been possible without the contributions of great people. Starting with Simpple S.L, in his representation Dr. Xavier Guardiola, which provided the prescription record database, guidance and great ideas. The guidance, interpretations and decisions provided by Dr. Jordi Duch and Dr. Robert Rallo have been essential as well.

I would like to mention Dr. Jordi Castellà Roca for giving general advises, which I always appreciate and Dr. Josep Domingo for sharing with me his opinions on privacy related concerns.

Finally, I wish to thank my parents for their support and encouragement throughout this thesis.

Bibliography

[1] Thomas E. Backer. Introduction. *Journal of Health Communication*, 10(4):285–288, 2005.

[2] Donald M Berwick. Disseminating innovations in health care. *Jama*, 289(15):1969–1975, 2003.

[3] Kenneth H Cohn and Douglas E Hough. The business of healthcare. 2008.

[4] James S Coleman, Elihu Katz, and Hebbekt Menzel. The diffusion of an innovation among physicians. 1957.

[5] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[6] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):21, 2012.

[7] Lawrence O Gostin. Marketing pharmaceuticals: a constitutional right to sell prescriber-identified data? *JAMA*, 307(8):787–788, 2012.

[8] Shin-Yi Chou Muzhe Yang, Hsien-Ming Lien. Is there a physician peer effect? evidence from new drug prescriptions. 2014.

[9] Harikesh Sasikumar Nair, Puneet Manchanda, and Tulikaa Bhatia. Asymmetric peer effects in physician prescription behavior: The role of opinion leaders. 2006.

[10] Everett M Rogers. Diffusion of innovations. 2010.

[11] MD Ryan M. Nunley and the Washington Health Policy Fellows. Habit-forming: Access to physician prescribing patterns, 2007.

[12] Thomas W. Valente and Rebecca L. Davis. Accelerating the diffusion of innovations using opinion leaders. *The Annals of the American Academy of Political and Social Science*, 566(The Social Diffusion of Ideas and Things):55–67, 1999.

[13] Thomas W Valente and Patchareeya Pumpuang. Identifying opinion leaders to promote behavior change. *Health Education & Behavior*, 34(6):881–896, 2007.

[14] James Vicini. Supreme court strikes down state drug data-mining, 2011.

[15] Andrew Zajac. A prescription for snooping, 2009.