Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 64 (2015) 483 – 490

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2015 October 7-9, 2015

Anonimity in health-oriented social networks

Andrei Vasilateanu*, Carmen Casaru

University Politehnica Bucharest, Splaiul Independentei nr. 313, Bucharest 060042, Romania

Abstract

Web 2.0 philosophy of content creation has spread to many domains including healthcare. A new trend in social networks is to serve specialized, niche-oriented participants, with unique interests. We study what are the advantages, disadvantages and concerns when deploying a social network in the health domain; to be used for sharing information about treatment outcomes and conditions. Our paper raises the privacy concern, namely anonymization, when different stakeholders use such a platform.

Peer-review under responsibility of SciKA - Association for Promotion and Dissemination of Scientific Knowledge

Keywords: e-health; social networks; anonymization; patient-empowerement.

1. Introduction

Healthcare worldwide is a system under increasing and complex pressures, mainly demographic and financial. The increase in life expectancy has shifted the focus from acute diseases towards chronic diseases, especially in developed countries. To give only one quantity as an example, the proportion of very old people (over 80 years old) will constitute 10% of the population in Europe in 2050 as opposed to only 3% today[1].

Reengineering of current healthcare systems is needed to support this new situation, with IT as an enabler and a driver for change. E-health, as a new domain, encompasses business, public healthcare and medical informatics, promotes new practices and paradigms, large-scale and small-scale, centralized and distributed.

* Corresponding author. Tel.: +40723167790; fax: +4021 402 9111.
E-mail address: andrei.vasilateanu@upb.ro

Peer-review under responsibility of SciKA - Association for Promotion and Dissemination of Scientific Knowledge
doi:10.1016/j.procs.2015.08.547

This paper contributes to a specific area of e-health, namely the use of social networks for patient empowerment and for the anonymous gathering of data by healthcare providers from such platforms. The structure of the paper is the following: we first describe the state-of-the art of using social networks in healthcare, we present different algorithms for anonymization of records, we present our prototype for an anonymous social network and then we draw the conclusions.

2. Healthcare-oriented social networks

Webster's dictionary defines health care as the "efforts made to maintain or restore health especially by trained and licensed professionals".

However, in nowadays' context, in the age of social networking and health data boom, the patients play an important role in the process of health care. They have become actively involved in the health care process by taking part into several areas as measuring, tracking, experimenting and engaging in interventions, treatments and research. Consumers are starting to do this individually, in collaboration with health peers, and in co-care with physicians and other medical professionals. "The entire health care process in general is starting to become more collaborative, moving to a co-diagnosis, co-care model between physicians, patients and other parties"[2].

Just as the role of the patient is evolving, so is that of the physician, by becoming a consultant and a creator of tailored care plans[3].

Peer-based health networks are bound to become a powerful member of the health care process with an expanding role, possibly having influence in policy, ethics, regulation, research and finance.[4]

Social networks have become an important tool for bringing people with shared interests together to interact. The following definition helps clarify what a health social network is and the main services that it has to offer:

"A health social network is a website where consumers may be able to find health resources at a number of different levels. Services may range from a basic tier of emotional support and information sharing to Q&A with physicians, to quantified self-tracking, to clinical trials access."[6] One of the key aspects that this type of networks has to offer is the possibility of finding other patients with similar health situations with whom one can share information about conditions, symptoms and treatments. Sharing this type of information helps patients learn more and gather more information with respect to their conditions. Thus, they will have a better overview of the situation, helping them with their decision making process.

Health social networks are primarily directed at patients, but caretakers, researchers and other interested and knowledgeable parties may be able to participate.

The largest and best-known health social network is PatientsLikeMe, which started in 2004 and had, as of January 2015, 300.000 members (http://www.patientslikeme.com).

2.1. Classification

Health social networks are mainly of two types. There are those that use as resource a large and wide variety of conditions (MDJunction, HealthChapter): this type of platforms are able to offer a more comprehensive look at a patient's health by covering a broader range of conditions than traditional medicine is usually able to offer. Secondly there are those that focus on fewer conditions in a more rigorous manner (PatientsLikeMe, CureTogether): this type of platforms offer additional services based on quantified self-tracking and collaborative filtering to identify potentially related conditions patients might be experiencing and match patients in similar situations.

Health social networks usually offer four main categories of services:

1. Emotional support and information sharing.

Patients have the option of interacting with others suffering from conditions similar to their own. This type of interactions helps them realize that "they are not alone", that there are others experiencing the same type of problems. The mere participation in such health care processes and platforms may get the patients feeling involved. It makes them part of a wide and varied community in which they can make a difference just by creating a profile and recording health information, seeing how other non-medical professionals describe the same conditions and symptoms and finding out what remedies others have tried. Last, but not least, and maybe one of the most important sources lays in the user interaction. Certain platforms allow and encourage direct interaction among users. Members are free to visit each other's profiles, send messages, leave comments and transmit lightweight social greetings. Thus, users can offer their advice and support to others, strengthening the bonds of community membership.

2. Ability to pose questions to physicians (MedHelp, WellSphere, MDJunction, iMedix, WeGoHealth).

This type of platforms usually have doctor profile pages, where the physicians add complete information regarding their expertise, background and affiliations along with corresponding previous Q&A sessions. Q&A sessions are usually displayed publicly (unless they are marked as private by the inquiring patient), so that each member has access to this piece of knowledge.

3. Quantified self-tracking (PatientsLikeMe, CureTogether, MedHelp, SugarStats).

The self-tracking functionality usually consists of easy to use data entry screens for conditions, symptoms, treatments and other related information. The information is then further processed and centralized into aggregated reports and statistical values. Usually this information is made available to the patient trough graphical display, by individual, aggregated population or custom groups.

4. Information regarding clinical trials.

"Even the presence of health social networks makes traditional clinical trials more efficient through the availability of large searchable inline databases of patients with health history and condition information"[5]. Pharmaceutical companies, industry analysts, policy architects and other interested parties can assess demand and market size directly from health social network platforms. For example, in May 2008, Novatis recruited clinical trial participants from PatientsLike ME, estimating that they were able to speed up their 1,200 patient study of a new medicine for Multiple Sclerosis by a few months. The health social network platforms provide rich sets of useful data generated through large online patient communities that interact and monitor their conditions. By processing and then further analyzing this data new findings can be developed to give a better understanding of the underlying conditions. The platforms also provide a feedback loop to the clinical trial process. Patients can become actively involved by providing suggestions and improvements to current functionalities. The conjunction between online health tracking with clinical trials allows patients to offer valuable experience feedback, including response to drugs and medicine.

2.2. Privacy issues

While there is no doubt that health social network platforms are a possible future solution for a better understanding of health conditions, their causes, their effects, the ways to handle, treat and finally prevent them, there are several aspects that may prove to be critical and need to be addressed[7].

One such aspect, that this paper is committed to cover, is related to the patients' privacy when their data is used.

One of the most useful characteristics of such platforms is the large volume of real-life patient generated data. This data is vital in the process of analysis for better understanding of the conditions' underlying causes, plausible treatments and prevention plans. Such information gathering and centralization would not be possible in the absence of the patients' involvement

It is true, that all such platforms have well defined and up to date user privacy and user agreement stipulation, but one must wonder, is it enough?

Patients have become more active and involved in the health care process. They are often willing to share their experiences and knowledge in order to interact, connect and help others which are going thorough similar experiences.

Laws like HIPAA and legal stipulations have been designed and put into place, in order to protect and harbor the identity of such platform patients. This type of legislation mostly requires that directly identifiable information such as the full name of the patient, the address, the phone number etc., be eliminated prior to publishing or handing over vital information with highly personal value.

Even so, there are cases in which the application of such stipulations is not enough to conceal and protect the members' identity. Even if data holders often remove explicit identifiers, identities can still be inferred by combining the released information. Current health social networks do not cover this issue which is a problem for their acceptance and trust.

In the next section we will present such a case of identity inferral and discuss different anonymization algorithms.

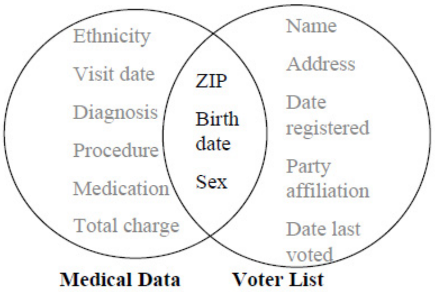


Fig. 1. Identification.

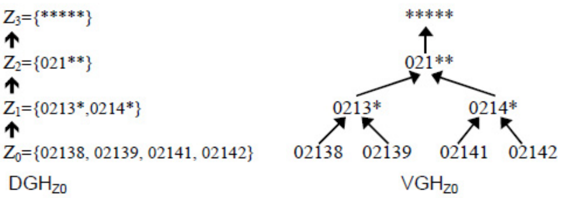


Fig. 2. Generalizing and suppressing ZIP codes (from [8]).

3. De-identification vs anonymization

As stated in the previous section, de-identification by removal of explicit identifiers (name, address, phone number, etc) has often proven insufficient in terms of identity protection and privacy.

3.1. Re-identification by Linking Example

Such situations have been known to occur in the case of re-identification by linking. One of these examples is what happened in USA where more states had to collect care data from health care providers. In Massachusetts, patients` specific data was collected containing information such as ethnicity, birth date, gender, ZIP code, visit date, diagnosis, procedure, medication, total charge. Since this data was thought to be anonymous it was given freely to researchers and industry. At the same time the voters list was available for purchase containing information as: name, address, ZIP code, birth date, gender. Aggregating the two subsets, as seen in [5]. allowed inferences to be made, for example it was possible to identify the governor of Massachusetts, since six people had his birthdate, three of them men and only one with the same ZIP code.

De-identifying information provides no guarantee of anonymity. "Given that a lot of sensitive information and valuable knowledge are hidden in data, the outsourcing of data is vulnerable to privacy crises and leads to demands for generalization or suppressing techniques to protect data from re-identification attacks"[8]

3.2. K-Anonymity

Anonymized data can be described by the k-anonymity property which helps protect released data against re-identification of individuals to which the data refers. With this mechanism, information for each person contained in the released data set cannot be distinguished from at least (k-1) individuals whose information also appears in the release. It provides privacy protection by guaranteeing that each released record will relate to at least k individuals, even if the records are directly linked to external information.

Having the private table, the released table RT, the attributes A1,A2..An and the quasi-identifier subsets Ai…Aj the definition is:

Let RT(A1,A2..An) be a table. QIRT=(Ai…Aj) be the quasi-identifier associated with RT, Ai…Aj ⊆ (A1,A2..An), and RT satisfy k-anonymity. Then, each sequence of values in RT[Ax] appears with at least k occurrences in RT[QIRT] for x=i..j

There are mainly two types of attributes:

- Key attributes: uniquely identify a particular individual (name, address, phone number, etc)
- Sensitive attributes: usually contain data which is vital for the research topic at hand (data related to medical records); thus they should not be further processes.

Quasi-identifiers represent the variables to be de-identified from the original dataset (dates, locations, race, ethnicity, gender, marital status, etc.) They can be used for linking anonymized data sets with other external data sets in an attempt of uncovering possible identities of various individuals.

Other anonymization forms exist, extensions of k-anonymity, such as l-diversity or t-closeness, but for the scope of this paper, we will relate only to k-anonymity.

Two methods exist for achieving k-anonymity: suppression of attributes and generalization, where a value is replaced by a less specific one. For example two ZIP codes {02138.02139} can be generalized to 0213*, indicating a larger geographical area. In a classical relational database system, domains are used to describe the set of values that attributes assume. For example, there might be a ZIP domain, a number domain and a string domain.

In order to achieve k-anonymity the ZIP codes need to be stripped down to a less informative state. So, there is a more general, less specific domain that can be used to describe ZIPs, say Z1, in which the last digit has been replaced by 0 (or removed altogether). There is also a mapping from Z0 to Z1, such as 02139 -> 0213*, as seen in Fig. 2.

We present next some algorithms for achieving k-anonymity.

3.3. K-Anonymity algorithms

In Datafly Algorithm, the data holder declares specific attributes and tuples in the original private table (PT) as being eligible for release. Next a subset of attributes of PT are grouped into one or more quasi-identifiers (QIi) and a weight from 0 to 1 is assigned to each attribute within each QIi that specifies the likelihood the attribute will be used for linking. The data holder specifies a minimum anonymity level that computes to a value for k.

Finally, a weight from 0 to 1 is assigned to each attribute within each QIi to state a preference of which attributes to distort[6].

The algorithm computes the frequency list for the values in PT where each sequence in the frequency represents one or more tuples in a table. While there exist frequencies occurring less than k-times, accounting for more than k-tuples the attribute with the most distinct values is generalized. In the end a table is returned in which the values stored as a sequence in frequency table appear as a tuple replicated to the stored frequency.

While Datafly solves k-anonymity, the results might not be k-minimal distortions, meaning unnecessary data loss occurs.

In Icognito Algorithm[9] the set of all possible k-anonymous full-domain generalizations of a table T are generated. Starting from single-attribute subsets of the quasi-identifier the algorithm iterates by checking k-anonymity in even larger subsets.

Mondrian Algorithm proposes a multidimensional recording model and a greedy algorithm for k-anonymization.

When applying anonymization to the healthcare domain we can start by stating the rules imposed by HIPAA (The Health Insurance Portability and Accountability Act) which was designed to protect the privacy of individually identifiable health information. HIPAA states that, when protected health information (PHI) is used for research purposes "covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information"[10] . The list of the 18 identifiers is the following: names, geographical subdivisions smaller than a state, relevant dates, phone and fax numbers, e-mails, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate/license numbers, vehicle identifiers and serial numbers, device identifiers and serial numbers, URLs, IPs, biometric identifiers, full face photographic images and any other unique identifying numbers.

4. MedipalCommunity prototype

In order to analyze the issues of health data privacy and protection against de-identification, we have designed and built a health social network platform prototype.

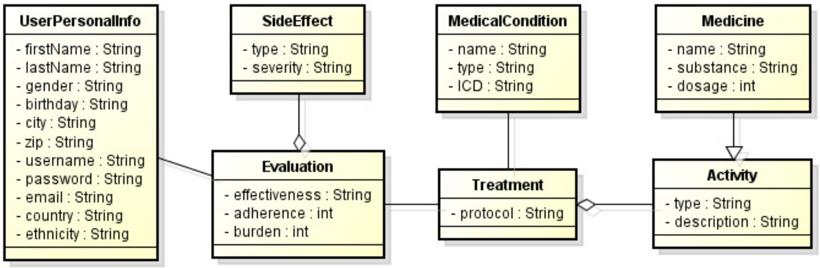


Fig. 3. MediPal Domain Model.

The aim of the study was to offer a complete solution between the standard services and functionalities that health social network platforms have to offer and a solution of effectively de-identifying and anonymizing patient personal information, without interfering with the quality of the sensitive data (data that is to be further processed, researched and analyzed).

The MediPalCommunity prototype was designed to offer services in three of the four available ranges of services:

- Emotional support and information sharing
- Quantified self-tracking
- Information regarding clinical trials

In this case, the platform serves as the main healthcare information collection tool. Through actions performed within the system, users provide valuable information that will be further stored within the platform's database, for future use. The data can then be exported from the database and anonymized. The prototype offers valid privacy and user agreement stipulations that the users can access prior and afterwards subscription. Also valid contact information is provided, so that users can keep in touch and communicate possible problems and suggestions in the prospect of offering a better experience and services.

The platform was intended for free patient use including free subscription and account creation. Thus, each patient will have his/her own custom profile with the corresponding personal and medical information. Patients will have access to treatments based on certain conditions groups and they will be able to provide feedback related to these treatments. The platform will also provide treatment recommendations based on real-time metrics computed using previous feedback received from other patients.

Finally, the health social network will also provide the option of exporting and anonymizing medical data that can be further delivered to research centers and pharmaceutical companies for processing and analysis.

The focus of our research in this paper is the collection and safe dissemination of data without compromising patient/user identity/privacy.

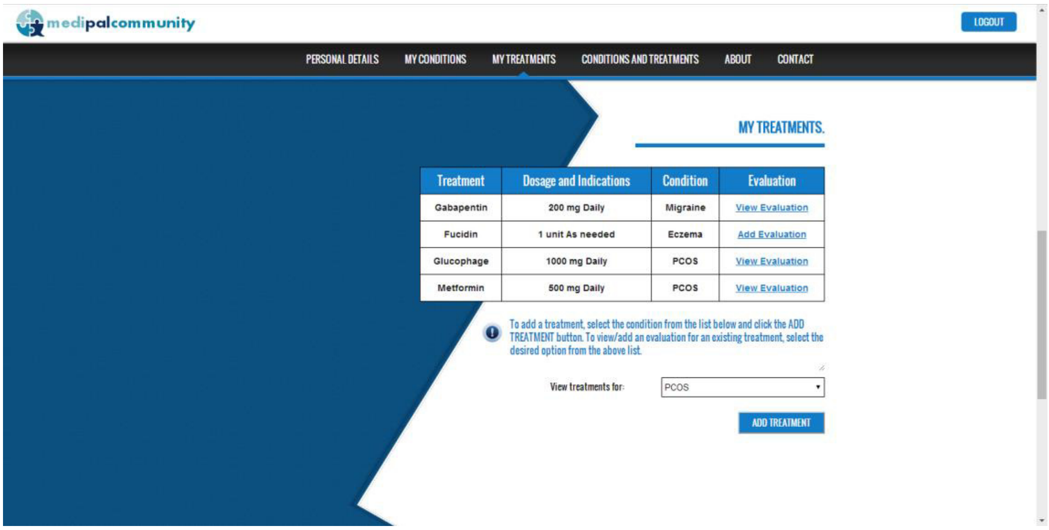
The platform, among other functionalities, offers support for exporting all or part of the healthcare information. in a format compatible with the UTD Anonymization Tool Box which is an open source toolbox put together by The UT Dallas Data Security and Privacy Lab, containing various anonymization methods for public use, in the benefit of researchers.

For the context of this experiment the following data attributes were chosen to be exported out of the full set of available information: {age, gender, country, condition, treatment, treatment effectiveness, treatment side effects, treatment adherence and treatment administration burden}.

One of the first activities that need to be carried out at the beginning of this process is related to a thorough analysis of the healthcare data. This analysis should result in a better understanding of the input data sets, together with a complete list of sensitive attributes and quasi – identifiers.

In this case, we can easily determine that the sensitive attributes are represented by the {medical information, treatment details, evaluation} subset. Thus, the settings and configuration of the tool box should avoid changing in any way this data, as any modification may lead to a decrease in data quality.

By further analyzing the input data set we can deduce that in some cases the city value may prove to offer too many details, when correlated together with the country and/or age and/or gender values. Due to this fact, it is recommended that the city value be removed from the original data set and all provisions that apply to directly linked values should be applied to this element values as well.



Treatment	Dosage and Indications	Condition	Evaluation
Gabapentin	200 mg Daily	Migraine	View Evaluation
Fucidin	1 unit As needed	Eczema	Add Evaluation
Glucophage	1000 mg Daily	PCOS	View Evaluation
Metformin	500 mg Daily	PCOS	View Evaluation

To add a treatment, select the condition from the list below and click the ADD TREATMENT button. To view/add an evaluation for an existing treatment, select the desired option from the above list.

View treatments for: PCOS

Fig. 4. MediPal Treatments Page.

Another problematic association may prove from correlating age and gender values. In some cases, this information alone may be used to infer the identity of patients. In the absence of the city values from within the context, further associations with country values may render useless.

Thus the anonymization should be applied to the minimal set of quasi-identifier elements which is comprised out of "age" and "gender" values.

After having determined the complete set of sensitive attributes and quasi – identifiers the configuration settings for the UTD Anonymization Tool Box need to be defined. The following details need to be provided:

- the data input file should point to the platform's exported data file,
- the output file should be set accordingly,
- the set of quasi-identifiers need to be defined, taking into account their corresponding indices within the exported data,
- the anonymization algorithm to be used and the anonymization threshold value (k).

For each quasi-identifier attribute, generalization ranges need to be defined at various levels based on the amount of input values and elements that can be accepted. While the "gender" element can only take two distinct values, the "age" element can accept a diverse range of discrete values. It is very important that each quasi-identifier use integer values, in order to allow the Tool Box to successfully process and analyze data. So, in the case of "gender" the "Female" value will replaced with 0 while 1 will correspond to "Male" values.

The Tool Box will process the data registered as quasi-identifier values and will try, after performing an inquiry based on this set of quasi-identifiers, to retrieve at least k similar entries, thus making it impossible to point towards one single person. If the number of retrieved entries is less than k, the Tool Box will begin replacing the original values with the VGH ranges defined at the previous step and resume testing. Thus, the tool will go towards the root of the hierarchy, until the minimum k threshold condition is fulfilled or until the very root of the VGH ranges is reached.

In this case only the values of "age" and "gender" will undergo suppression and/or generalization. For example a 56 age value will at the first step be replaced with the [50, 100) range of values.

5. Conclusions

Although, the currently available health social network platforms offer classical and standard data protection and security solutions like authentication and authorization, they do not offer complete and efficient solutions for protection against de-identification. Most such platforms comply with legal stipulations of removing directly identifiable information prior to publishing or handing over vital information with highly personal value. But as analyzed, discussed, argued and proven in this paper and in larger scope, such actions are not enough to ensure the safe-keeping of patient privacy.

The actions and interactions of patients in such social platforms, their feedback and involvement are becoming more and more vital throughout this process of health care evolution and growth.

This is why it is very important that patients who are sharing their information and data be able to trust the platform they are working with, without having to worry about their identity becoming unraveled.

In conclusion, health care social network platforms need to concern themselves more with the safe-keeping and protection of patient privacy and identity. The data should be carefully analyzed and a complete anonymization and de-identification process should be available and executed before releasing sensitive data.

Of course, as the way in which data is gathered and processed evolves, a need for new optimized anonymization solutions increases, rendering the classic solutions outdated.

References

1. Commision of the European Communities. (2002). Europe`s response to World Ageing Promoting economic and social progress in an ageing world.
2. Brophy-Warren J. The New Examined Life: Why more people are spilling the statistics of their lives on the Web. *The Wall Street Journal*, 2008.
3. Charles C, Whelan T, Gafni A. What do we mean by partnership in making decisions about treatment?. *BMJ: British Medical Journal* 1999; 319.
4. Wuyts, K., Verhenneman, G., Scandariato, R., Joosen, W., & Dumortier, J. What electronic health records don't know just yet. A privacy analysis for patient communities and health records interaction. *Health and Technology* 2012; **2:**159-183.
5. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *International journal of environmental research and public health* 2009; **2**:492-525.
6. Hsu CH, Tsai HP. KAMP: Preserving k-Anonymity for Combinations of Patterns. In: *Proceedings of IEEE 14th International Conference on Mobile Data Management*; 2013, p. 97.
7. Lee, Newton. *Facebook nation: Total information awareness*. Springer, 2014.
8. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002; 571-588.
9. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*.
10. Centers For Disease Control and Prevention. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. MMWR: Morbidity and Mortality Weekly Report 2003; **52**:1-17 .
11. Casaru C. Data Protection and Privacy K-Anonymity. Universitatea Politehnica Bucuresti Romania; 2014.