GraPhys: Understanding Health Care Insurance Data through Graph Analytics

Luis G. Moyano, Ana Paula Appel, Vagner F. de Santana,
Marcia Ito, Thiago D. dos Santos
IBM Research, Brazil
{lmoyano, apappel, vagsant, marciaito, thiagodo}@br.ibm.com

ABSTRACT

Healthcare insurance data represent a rich source of information and has the potential to contribute significantly in guiding business decision making. In this work we present GraPhys, a Graph Analysis platform designed for exploration, visualization and analysis of healthcare insurance data and its corresponding metadata. By taking advantage of relationships contained in healthcare claims data, we are able to apply Graph Analytics methods and algorithms in order to devise useful business metrics to guide data analysis and exploration. Our tool focuses in better understanding physicians, patients and their practices. We illustrate our approach by demonstrating two use cases where we show how graph analytics metrics, combined with other data, may lead to useful insights not directly available to traditional Business Analytics.

Keywords

graph mining, health, visual mining

1. INTRODUCTION

Healthcare insurance companies generate large amounts of data as a product of their daily operations. These vast amount of data reflects the complexity in their operations, as it demands precise tracking of multiple pieces of information mainly associated to accurately account for costs and expenses of all medical services provided by healthcare professionals to patients. These transactional data are very rich, as it contains spatial temporal coordinates as well as information about the parties involved, plus many more variables, from disease codes to diagnostics in the form of unstructured, free text.

The information contained in these data can be critical for decision making on several levels, from making better medical procedures to improving business processes. Indeed, it is easy to realize that such rich source of data may contain clues to better understanding patterns in the behavior of doctors, patients, healthcare providers, and other stakeholders. This has been the final objective of whole areas such as Business Intelligence and Business Analytics [2].

Traditional Business Analytics regards data querying and aggregation as the main building blocks of information processing. While this approach is fruitful and necessary, there are data structures that contain additional useful information: relationship data in the form of metadata. By relationship data we mean data structures that relate two or more entities in a precise, quantifiable way, such as a phone call connecting two people or two computers connected by a physical link. This is the case in healthcare insurance data, where patients relate to healthcare professionals by making consultation visits, performing exams, among other medical procedures.

This relationship metadata may be presented in the form of a graph, where nodes correspond to patients and healthcare professionals (in this case nodes are not all of the same type), and links corresponds to any medical procedure that has involved a given pair of patient and doctor [10, 4]. Once it has been mapped to a graph structure, graph analytics techniques can be applied in order to better understand patterns in the data and extract useful insights for business decision making [9, 7, 3, 6, 5].

The main contribution of this work is the provision of a platform for graph analytics in healthcare insurance claims data, allowing for better visualization, exploration and analysis of the data. Our goal with this platform is to take advantage of the relationship data in a systematic way, exploring graph analytical techniques to better understand doctors, patients and other stakeholders as well as their practices.

Our platform has been tested with real claims data from a major Brazilian healthcare insurance company. The data corresponds to more than 18 months of claims transactions nation-wide, totaling more than 2.1 million patients and more than 220.000 doctors and medical professionals.

2. RELATIONSHIP EXPLORATION AND VISUALIZATION

Health care insurance companies and other medical insurance providers have a wealth of data at their disposal. Operations in this area produces an amount of information, specially transactional, and also other, more persistent, forms of data, such as demographics, healthcare provider location, and other essential data to accurately carry forward all business processes.

WWW '16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890544

An important piece of transactional data are claims, which for the present work may be regarded as a report from the physician or healthcare provider to the insurance company. A claim informs all the details of a patient's visit or medical procedure. Even though claims data may vary, it generally contains an ID of the healthcare professional involved in the procedure (it may also be a group of professionals), and ID of the patient who was treated and a timestamp corresponding to the moment the event took place. Further vital information is usually added, such as which types of health services were delivered, and the associated costs owed for the insurance company to process, among others.

Claims data may be mapped in graph form by considering the healthcare professional as a node (of a certain type), patients also as nodes (of a different type), and establishing a link whenever there is a claim containing both. The resulting graph is an example of *bipartite graph*, where there are several types of nodes and there are no links between nodes of the same type. Moreover, additional bits of information in the data may be included in the graph in the form of weights in the links (such as the timestamp, the service provided or the expense amount), or as attributes in the nodes (as the patient demographics information or the medical specialty of the healthcare professional). Both weights and attributes may be represented in a graph layout by mapping their values to colors or sizes in the case of nodes, or colors and widths in the case of links.

Up to now, we have described a mapping of claims data into graph form, which is particularly adept to being represented visually for exploration. Additionally, the GraphPhys platform contains additional metrics, computed internally, designed to explore other aspects of interest in the healthcare claims data. These metrics were designed specifically taking into consideration the interconnected nature of the data, exploiting the possibilities of graph analysis. The new metrics can be combined with the original data contained in the dataset. The combination of both sources of information may yield improved analysis and alternative exploration patterns, potentially leading to new insights. We will discuss specific examples in Secs. 4.1.1 and 4.1.2.

3. ARCHITECTURE

The GraphPhys platform was develop using IBM Bluemix platform (Platform as a Service - PaaS)[1], IBM's open cloud platform that provides mobile and Web developers access to IBM software for integration, security, transaction, and other key functions, as well as software from business partners.

In Figure 1 we can see the different components that cover our platform, including a Data Storage component, an Application, a Graph layout component, a Histogram component, and a Visualization component.

Data source. Data are extracted from a designated source file and loaded to a JSON file, structured as nodes and links. The data are saved to the database platform through a javascript application.

Data Storage. We chose Cloudant DB [2] [8] as the database platform to store all healthcare claims data, as it is compatible with IBM Softlayer, the cloud platform utilized for the development of the demo. There are two databases

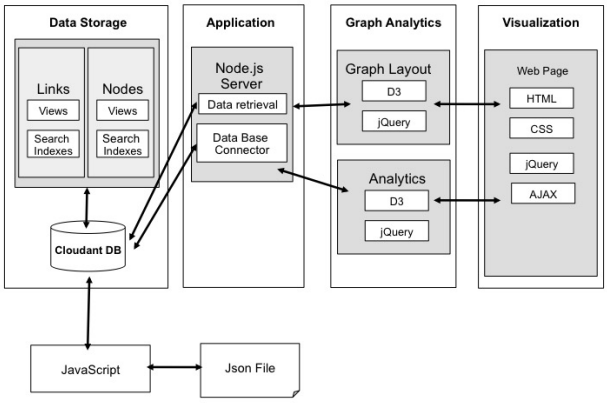

Figure 1: GraPhys platform architecture.

involves, one for links and another one for nodes. Each one of these database has views and search indexes with a description of all methods through which the application gets the data.

Application Node.js Server (App.js). This component is responsible for the server description, as well as the authentication configuration parameters to access the Cloudant Database. Queries to the database solicited by the web page are done through this component.

Graph Analytics. The Graph Analytics contains the Graph Layout submodule and the Analytics submodule. The Graph Layout module uses the D3 library to render the graph layout. When receiving the web page request with the filters chosen by the user, the data is obtained from the database through a request to the server. With the data received to render the graph layout as a starting point, the Analytics component computes all necessary metrics, statistics and distributions (for instance, degree and other centrality measures), using the D3 library.

Visualization. The platform has been implemented as an HTML web page with CSS design. It uses jQuery and Ajax to interact with the user, to make HTTP requests and to update content. The graph is designed using the D3 library, through data obtained by requests to the database, and filtered according to the different user selections.

4. DEMO USE CASES

Conference participants will be able to interact directly with the GraphPhys platform. There will be a test dataset containing anonymized real data to explore different visualizations, layers of aggregation as well as filtering options. The test dataset is derived from real healthcare claims data from a major Brazilian healthcare insurance company. The demonstration will contain all the main components of the platform, as described in Sec. 3.

As first use case, we showcase two specific graph analytics metrics, *retention* and *reciprocity*, which we will describe in Section 4.1.1 and 4.1.2. We describe how these metrics fit in the exploration capabilities of the platform.

As a second use case , we describe the possibility of exploring and visualizing, always from a graph perspective, different aggregation alternative in the data, explained in

[1]http://bluemix.net/
[2]https://cloudant.com

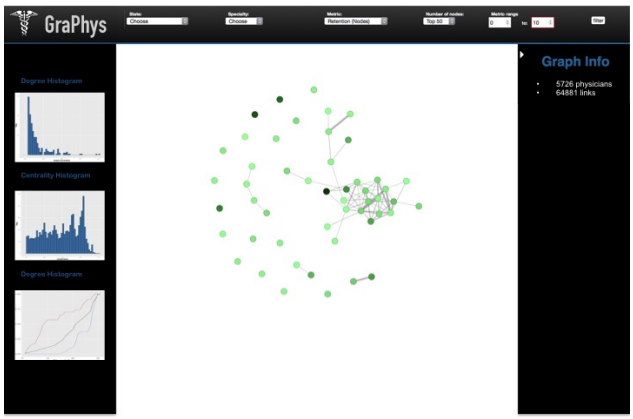

Figure 2: Front-end interface of the GraPhys platform. The use case shows a doctor-doctor graph, where nodes are color-coded proportionally to the *retention* metric.

more detail in Sec. 4.2.

4.1 Graph Analytics metrics

4.1.1 Retention

The first example that will be presented is related to the concept of *retention*, which aims at capturing the ability of a given healthcare professional to retain a patient who visits her with some frequency.

We compute the number of links (visits) of a given patient to all visited doctors, normalizing to obtain a relative measure. Combining the relative frequency of visits with the degree (i.e., the absolute number of visits) of each patient, we can filter and choose those patients that for some reason or another have a preferred doctor. Finally, we quantify how many of such patients any doctor has, allowing us to find those physicians that present an above-average number of such patients, which is a strong indication of a higher capacity of retaining patients.

This metric may be an indirect proxy for patient loyalty and allows, through the use of the tool and in combination with other data, to pinpoint interesting cases for the user of the platform to a more extended analysis.

In Figure 2 we can see our platform illustrating a doctor-doctor graph, i.e. all nodes are of the same type. The central layout renders the graph structured data, and nodes connected to each other correspond to doctors with patients in common. The color of the nodes is proportional to the *retention* variable previous explained, which is selected by the user from the drop-down menu in the top of the interface. In this setting, it is straightforward to see which are the doctors with the highest value of the metric, and how they relate to each other. Further information about the resulting graph can be found at both sides of the central layout, from the total number of nodes and links in the graph, to different histograms and figures that characterize different metrics of interest for the specific subset of doctors.

4.1.2 Reciprocity

In many cases, doctors and other healthcare professionals interact directly or indirectly with other members of the healthcare system. A doctor may recommend another physician or healthcare provider to initiate or continue treatment, or even refer the patient for treatment with doctors from other specialties are a better fit. This practice can be part of the normal workflow patients undergo and, depending on many factors, may occur with more or less frequency. On the other hand, there could be the case where two or more doctors share patients with frequencies off the typical rates found in similar cases [1].

These cases are interesting for the insurance company to understand further, and check if the atypical relationship is indeed justified or a case that should be rectified.

Our platform computes a special dedicated metric, which we call *reciprocity metric* in order to quantify and detect this kind of situation. The reciprocity metric is built over the concept of mutual degree, i.e. the sum of the number of directed links shared by two given nodes in the two possible ways, back (w_{ij}) and forth (w_{ji}). We penalize this sum by subtracting the absolute value of the difference ($|w_{ij} - w_{ji}|$), in order to reveal asymmetries in both directions. Properly normalized, this metric captures pairs of doctors that share an unusual large number of patients, in particular the case where this is true in both directions.

4.2 Aggregated Graph Data Exploration

Healthcare data, as most types of data, has a layered structure which is essential to be taken into account into any analysis. These layers of information allow for various ways of data aggregation, which in turn leads to useful information and insights. For example, is to consider aggregating data by specialties, providing information about, for instance, all cardiologists, or all physicians that work in a given hospital, and so on.

Our graph analytics approach naturally combines graph data structures with any kind of variable aggregation, by transforming the graph appropriately. To illustrate this feature we may start considering claims in a given time window. As described previous, doctors and patients are mapped into nodes (yielding a bipartite graph or network) and links represent claims, which in turn stand for visits, exams, hospitalization and other medical procedures. We could now want to visualize a more summarized version of the same data by aggregating doctors by healthcare provider (hospital, clinic, first aid post, etc.). This is achieved in our platform by selecting the appropriate variable for aggregation, in this case "provider", which replaces individual doctor nodes by provider nodes, and rewires links appropriately. Each node will thus represent a healthcare provider, also standing for all doctors that work in that provider. To include this information, the tool has the option to resize nodes proportionally to the number of doctors associated to the node. There could be the case of a patient visiting many doctors in a hospital. In this case, links can also account vary their width proportionally to the number of claims.

We could now want to explore these data in an even higher level scale, say by specialty, maintaining its graph form. The user selects the appropriate tool function to aggregate both doctors and patients by the "specialty" variable, and repeat the procedure described previous, resulting in another graph.

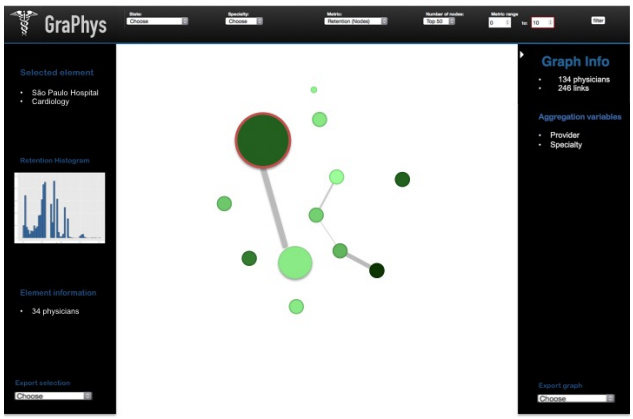

Figure 3: Front-end interface of the GraPhys platform. The use case shows a doctor-doctor graph aggregated by provider and specialty. Node size and link width are resized as explained in the text.

This second use case is described in Figure 3. We show a similar case as before, but doctors have been aggregated both by provider and specialty. Nodes and links have been resized as explained before. The user may also select a node or link, which is highlighted while related information is collected in the side panel, including statistics of the elements contained in the aggregated node or link.

5. CONCLUSION AND FUTURE WORK

We have presented our Graph Analytics platform for health insurance claims data. The tool was designed for exploration, visualization and analysis of claims data, leveraging the relationship metadata contained in the original data, by means of graph analytics. This data structure allows to naturally define new metrics based in graph theory algorithms and concepts which may be explored on its own, as well as in combination with other variables in the dataset.

The combination of graph analysis and exploration opens up a wide range of applications, for instance, rapidly detecting pairs of doctors that share patients in abnormal levels, or ranking doctors with the highest level of patient loyalty compared to same specialty, among many other examples.

GraPhys is evolving quickly in many fronts. One particular extension we are working on is adding new graph-theoretical metrics that address other aspects of the healthcare business. There are additional challenges ahead to achieve scalability for large (tens and hundreds of thousands of nodes) and very large datasets (millions of nodes). Finally, new functionalities from the analytics and reporting capabilities are being considered as further improvements.

6. REFERENCES

[1] A. Beutel, L. Akoglu, and C. Faloutsos. Fraud detection through graph-based user behavior modeling. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1696–1697. ACM, 2015.

[2] H. Chen, R. H. Chiang, and V. C. Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4):1165–1188, 2012.

[3] A. Chmiel, P. Klimek, and S. Thurner. Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, 16(11):115013, 2014.

[4] P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[5] P. Klimek, A. Kautzky-Willer, A. Chmiel, I. Schiller-Früwirth, and S. Thurner. Quantifying age-and gender-related diabetes comorbidity risks using nation-wide big claims data. *arXiv preprint arXiv:1310.7505*, 2013.

[6] B. E. Landon, J.-P. Onnela, N. L. Keating, M. L. Barnett, S. Paul, A. J. OâĂŹMalley, T. Keegan, and N. A. Christakis. Using administrative data to identify naturally occurring networks of physicians. *Medical care*, 51(8):715, 2013.

[7] K. D. Mandl, K. L. Olson, D. Mines, C. Liu, and F. Tian. Provider collaboration: cohesion, constellations, and shared patients. *Journal of general internal medicine*, 29(11):1499–1505, 2014.

[8] G. Mone. Beyond hadoop. *Communications of the ACM*, 56(1):22–24, 2013.

[9] S. K. Sauter, L. M. Neuhofer, G. Endel, P. Klimek, and G. Duftschmid. Analyzing healthcare provider centric networks through secondary use of health claims data. In *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*, pages 522–525. IEEE, 2014.

[10] F. Wang, U. Srinivasan, S. Uddin, and S. Chawla. Application of network analysis on healthcare. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 596–603. IEEE, 2014.