Journal of Biomedical Informatics 61 (2016) 203–213

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Using network projections to explore co-incidence and context in large clinical datasets: Application to homelessness among U.S. Veterans

Warren B.P. Pettey [a,b], Damon J.A. Toth [a,b,c], Andrew Redd [a,b], Marjorie E. Carter [a,b], Matthew H. Samore [a,b,d], Adi V. Gundlapalli [a,b,d,*]

[a] IDEAS Center, VA Salt Lake City Health Care System, Salt Lake City, UT, United States
[b] Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States
[c] Department Mathematics, University of Utah, Salt Lake City, UT, United States
[d] Department Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

ARTICLE INFO

*Article history:*
Received 30 April 2015
Revised 14 March 2016
Accepted 28 March 2016
Available online 1 April 2016

*Keywords:*
Homelessness
U.S. Veterans
Social network analysis
Information foraging
Data exploration
ICD9 codes

ABSTRACT

*Introduction:* Network projections of data can provide an efficient format for data exploration of co-incidence in large clinical datasets. We present and explore the utility of a network projection approach to finding patterns in health care data that could be exploited to prevent homelessness among U.S. Veterans.

*Method:* We divided Veteran ICD-9-CM (ICD9) data into two time periods (0–59 and 60–364 days prior to the first evidence of homelessness) and then used Pajek social network analysis software to visualize these data as three different networks. A *multi-relational network* simultaneously displayed the magnitude of ties between the most frequent ICD9 pairings. A *new association network* visualized ICD9 pairings that greatly increased or decreased. A *signed, subtraction network* visualized the presence, absence, and magnitude difference between ICD9 associations by time period.

*Result:* A cohort of 9468 U.S. Veterans was identified as having administrative evidence of homelessness and visits in both time periods. They were seen in 222,599 outpatient visits that generated 484,339 ICD9 codes (average of 11.4 (range 1–23) visits and 2.2 (range 1–60) ICD9 codes per visit). Using the three network projection methods, we were able to show distinct differences in the pattern of co-morbidities in the two time periods. In the more distant time period preceding homelessness, the network was dominated by routine health maintenance visits and physical ailment diagnoses. In the 59 days immediately prior to the homelessness identification, alcohol related diagnoses along with economic circumstances such as unemployment, legal circumstances, along with housing instability were noted.

*Conclusion:* Network visualizations of large clinical datasets traditionally treated as tabular and difficult to manipulate reveal rich, previously hidden connections between data variables related to homelessness. A key feature is the ability to visualize changes in variables with temporality and in proximity to the event of interest. These visualizations lend support to cognitive tasks such as exploration of large clinical datasets as a prelude to hypothesis generation.

Published by Elsevier Inc.

## 1. Introduction

Differential diagnosis, the process of thoughtfully using multiple evidence streams to identify the most likely of several plausible candidate causes, not only requires knowledge of events and their co-incidence, but also the context associated with those events. Consider, for example, how healthcare providers diagnose and manage patients by integrating and synthesizing patient reported complaints, history of illness, associated psychosocial factors, and results of laboratory tests and imaging studies. Sometimes one of these factors is, by itself, sufficient to produce an outcome (for example, loss of employment leading to homelessness). However, it is also the case that several smaller factors can combine as a viable precursor for an outcome (for example, a poorly fitting prosthetic, back pain, and joint pain leading to loss of employment) or become the actual cause (behavioral issues and alcohol abuse leading to eviction). When this process is scaled to population health levels, appreciating and investigating co-incidence and contexts—especially processes that operate at a smaller, local

* Corresponding author at: VA Salt Lake City Health Care System, 500 Foothill Drive (182), Salt Lake City, UT 84148, United States.
*E-mail address:* adi.gundlapalli@hsc.utah.edu (A.V. Gundlapalli).

http://dx.doi.org/10.1016/j.jbi.2016.03.023
1532-0464/Published by Elsevier Inc.

level—becomes computationally and cognitively challenging due to the magnitude fold increase in complexity of the variables.

In this paper, we present a method for representing and surveying large clinical datasets as network projections, facilitating the exploration of concept contexts through literal connections. Visually detecting and interpreting patterns in network presentations of data (a shared goal with *exploratory social network analysis*, which starts with, "no specific hypotheses beforehand that we can test" [1]) is a preliminary step in the continuum of the scientific method that could lead to hypothesis generation and planning for further traditional analyses. It is not a replacement for traditional statistical approaches, but rather a means to display expected relationships between medical and social conditions and also to display collections of conditions, "that hint at where additional information is needed and help to direct our attention to domains that need further investigation" [2].

The popularity of social network analysis (SNA) as a discipline and as a framework for analysis has grown rapidly alongside the availability of increasingly powerful computers. Part of this stems from the large improvements in processing power required for SNA's computationally intensive analytical methods, but also from the ability of computers to visualize networks on screen. The benefits of on-screen display are many, and include the ability (depending on the software and data) to rotate and move a network, to zoom into specific parts of a network, to apply various SNA metrics that describe parts of the network by connectedness, similarities, and differences, and to change the way a network appears to make certain features easier to view. Some software permits a three-dimensional drawing of a network, and some software will draw and even animate networks according to changes across time.

Many disciplines and groups have turned to SNA to describe, understand, and represent a variety of systems, objects, concepts, and more. Examples range from the serious (mapping networks of terrorist cells [3]) to the whimsical ("Diss Songs" in rap music [4]), extending from biological systems, sociological processes, and professional interactions to marketing, national security, politics, infrastructure, language sciences, and health. SNA analyses have included physician behavior, obesity, smoking, disease transmission, disease immunity, healthcare worker and patient interaction, vaccination strategies, substance abuse, suicide, and more [2,5–9]. A text on the topic of SNA and health concluded that, "Every health topic can be viewed through the network perspective" [8].

In a social network, concepts of interest (such as medical codes) are represented as vertices and the context (or co-incidence of these concepts) is represented as edges or lines between the vertices. Generally, colors, shapes, and sizes of vertices and edges represent attributes of interest or are the results of various analysis measures. Edge thickness and color can, respectively, represent the strength and type of the relation between vertices. As compared to tables, models, and description, networks allow, "several pieces of numerical information simultaneously, whereas numbers, mathematical formulae, and written language have to be read sequentially" [2].

Drawing a network requires a network projection or network layout algorithm. Several disciplines, including the social sciences, mathematics, physics, statistics, network science, and the information visualization community—each with its own motivations, needs, vocabularies, and traditions—have contributed network layout algorithms (we recommend Chapter 37: "Network Visualization" in [2] for an extensive review of network projections). Network layout algorithms draw a network's vertices and edges on a two-dimensional or three-dimensional space, balancing distortion and aesthetics. (It is a challenge familiar to cartographers, who must flatten (or project) a higher-dimension space against a lower-dimension space, attempting to minimize distortion for that aspect of the map that is most important.) Most SNA software comes prepackaged with many network layout algorithms, and some (such as Cytoscape and Gephi) have the ability to import additional algorithms.

Correctly inferring *similarity* from vertex-to-vertex proximity requires knowing how the projection method calculated vertex positions. Force-directed network projections (also called spring-embedded, or energized projections) balance a preference for uniformly drawing similarly weighted edges against various aesthetics rules that make the network easier to view (examples include Kamada-Kawai and Fruchterman Reingold). These networks tend to be aesthetically pleasing, intuitive, and are by far the most common drawing method. Interpreting vertex-vertex proximity in a force-directed network requires some caution, since spatial closeness is at least partially accidental. In contrast, there are a set of network projection algorithms that consider similarly connected vertices to be similar to one another, and therefore draw them closer together (examples include multi-dimensional scaling, Louvain and Visualization of Similarities (or VOS)). Many of these latter projection algorithms have complimentary community finding algorithms that partition or color vertices according to similar or identical rules, making it possible to draw a network using the aesthetically pleasing force-directed network projection algorithms and visualize similarity with community finding partitions.

Visualizing data and context as network projections decreases the physical footprint on screen and in print. It displays existing relationships with links and the absence of relationships with nothing, as compared to several other forms of graphical data presentation such co-incidence matrices, which use the same physical footprint for existing and non-existing relationships. (While this is not a problem for small data, it can approach unworkableness as data size scales up, with many variable values or when data are temporal.) Displaying the absence of these connections can also be important for understanding and planning.

We present three projection based methods that allow us to compare and contrast differences in medical codes for two different time periods: First, according to medical codes of interest; Second, according to a threshold of co-incidence between medical codes; Third, by a comparison of two networks. These three approaches let us consider the change in co-incidence between the most popular medical codes, the difference in the most co-incident medical codes, and a highlight of co-incident medical codes that spike and plummet immediately before homelessness.

Our use case for this study is that of investigating the co-incidence of medical codes for recent U.S. war Veterans who became homeless in an effort to identify points of intervention to prevent homelessness. Homelessness among Veterans is a matter of national importance. As compared to the overall U.S. population, Veterans are over-represented among the homeless. It is estimated that more than 136,000 Veterans spent at least one night in an emergency shelter or transitional housing program between October 1, 2008 and September 30, 2009 [10]. The U.S. Department of Veterans Affairs (VA) has been actively screening for homelessness among Veterans visiting VA medical facilities [11,12] along with community outreach to identify and serve homeless Veterans. Of course, it is preferable to find at-risk Veterans before homelessness occurs. With the availability of large, VA clinical datasets, it becomes possible to include this as a resource to support early identification of those Veterans at risk of homelessness by mining those data. While visual analytics and SNA techniques have been used to detect clinical event patterns using EHR data [13] and also to analyze populations of persons experiencing homelessness [14–17], they have not yet been used to explore the co-incidence of diagnosis codes assigned to homeless persons seeking medical care.