Network-Centric Data Mining Solutions for Prospective Medical Applications

Darcy A. Davis
Advisor: Nitesh V. Chawla
University of Notre Dame

Combined Genetic and Phenotypic Disease-Disease Networks

Due to common genetic, molecular, and environmental risk factors, many diseases are co-morbid, expressed in the same patient. Also, many common, chronic, and devastating diseases, such as cancer and diabetes, are complex diseases influenced by a combination of environment and epistasis between many genes. We use patient medical histories (phenotype data) and previously discovered disease-gene associations to construct, analyze, and compare disease-disease networks. Also, we merge the data into a multi-relational network to study the question:

What are the patterns of interplay between patients, diseases, and genes?

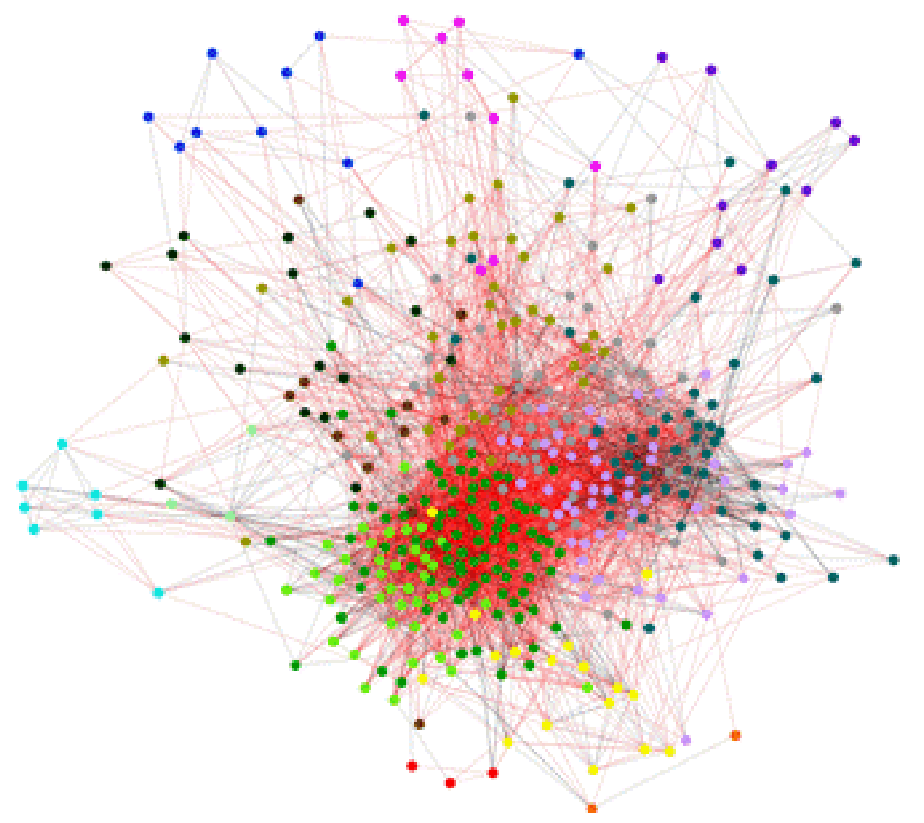
We focus on topological tools and interaction substructures. Understanding these building blocks and their probabilistic traits may be applicable to discovering disease-gene candidates or other unique interactions of interest.

Phenotypic Network

- Nodes are diseases
- Edges indicate that the diseases are co-morbid significantly more than randomly expected
- Edges weighted by mutual information

$$w(d_1,d_2)=\left(\frac{p(d_1,d_2)}{p(d_1)p(d_2)}\right)$$

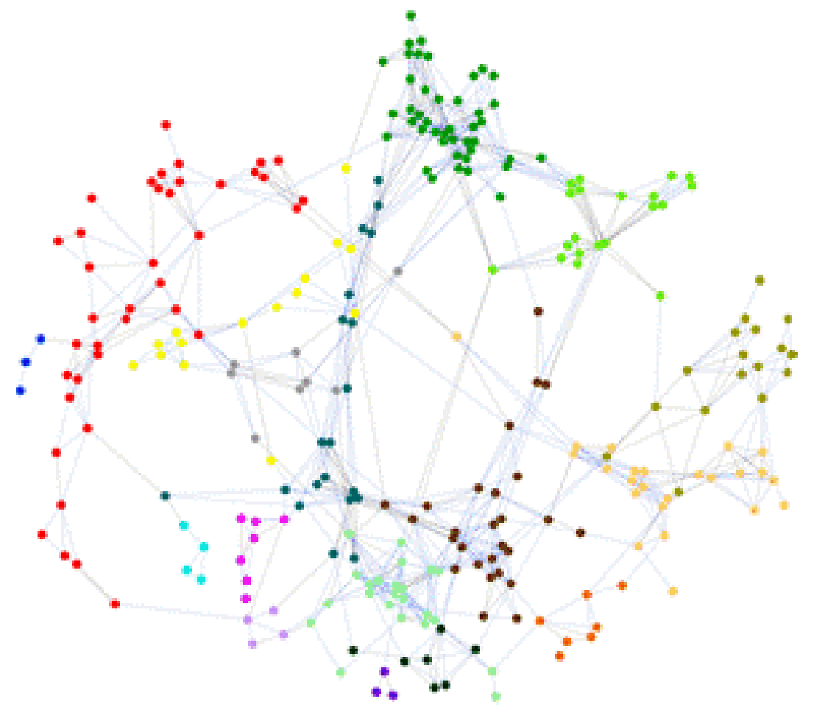
- Clustered using Walktrap



Phenotypic co-morbidities form a chaotic network with imprecise clusters

Genetic Network

- Nodes are same diseases represented in the phenotypic network
- Edges indicate that the diseases share gene associations significantly more than randomly expected
- Also weighted with mutual information, clustered with Walktrap



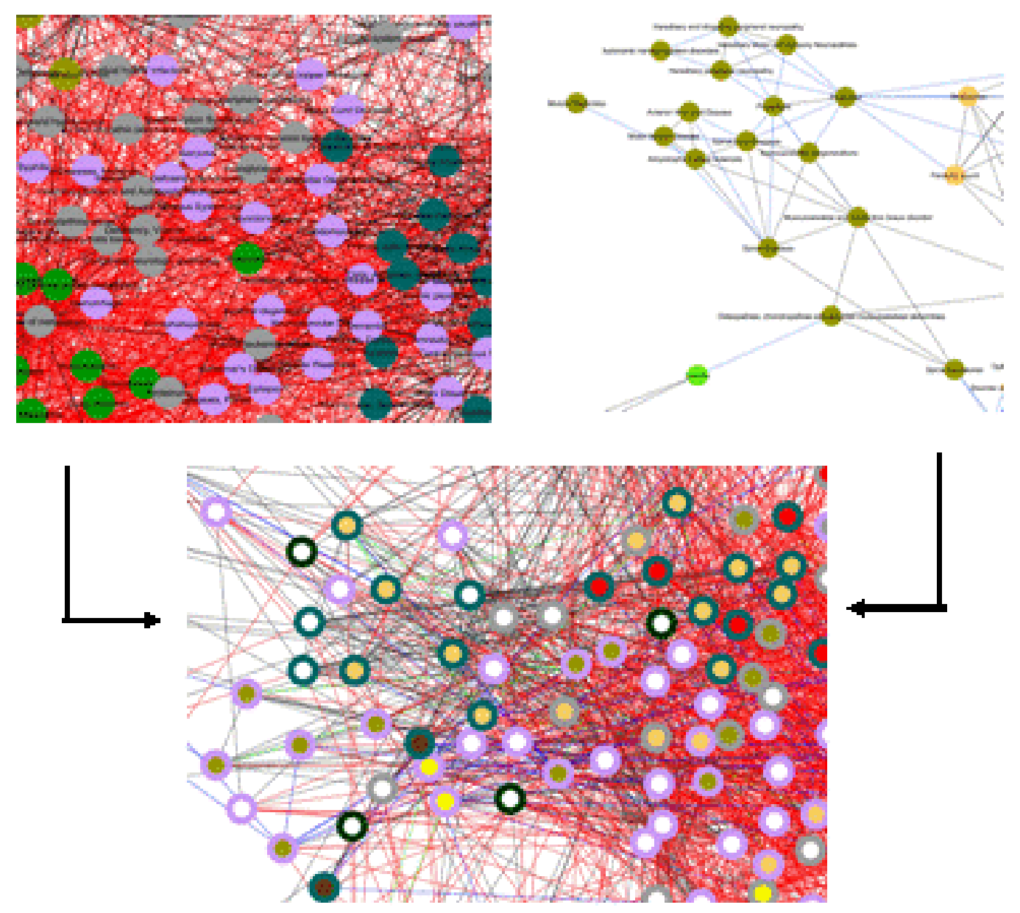
Genetic associations form clean structures with clear biological themes

Combined Multi-Relational Network

The two networks contain the same disease nodes with different patterns of connections and weights. Thus, they can overlaid into a single multi-relational network with multiple edge types. The original clusters can be preserved as node attributes.

Topological Tools and Link Prediction

We plan to study the substructure of our heterogeneous network using multi-relational expansions of network motifs and induced graphlet census. We will use these substructures and their associated probabilities to develop a link prediction method for multi-relational networks which also accounts for interplay between edge types.



Despite drastic structural differences, there is significant overlap between phenotypic an genetic clusters in the merged network.

Introduction

Faced with unsustainable costs and enormous amounts of under-utilized data, health care needs more efficient practices, research, and tools to harness the benefits of data. These methods should create a feedback loop where computational tools guide and facilitate research, leading to improved biological knowledge and clinical standards, which in turn should generate better data. In order to facilitate the necessary changes, better tools are needed for assessing risk and optimizing treatments, which further require better understanding of disease interdependencies, genetic influence, and translation into a patient's future. We propose network-centric data mining approaches for benefit in multiple stages of this feedback loop: for better understanding of disease mechanisms and for development novel clinical tools for personalized and prospective medicine.

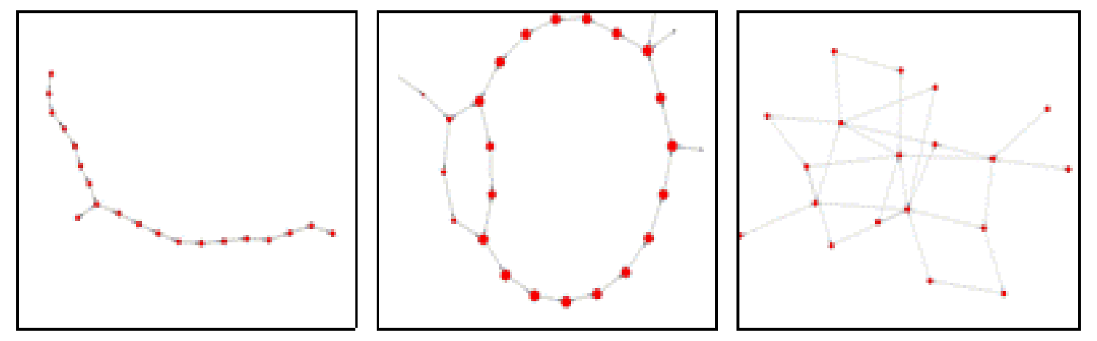
Network Analysis of NICU Team Structure

Patients in the neonatal intensive care unit (NICU) are at relatively high risk for preventable medical harm. Long and complex stays, with up to 300 nursing handoffs, leave infants at risk. Maintaining an well functioning nursing team is a daunting challenge. Clinical studies have shown organizational characteristics of care to be predictors of performance, and networks have been used to study performance and fault tolerance with relation to team structure in other domains. We demonstrate of the use of network analysis to answer the question:

How does the structure of the nursing team affect the quality of care?

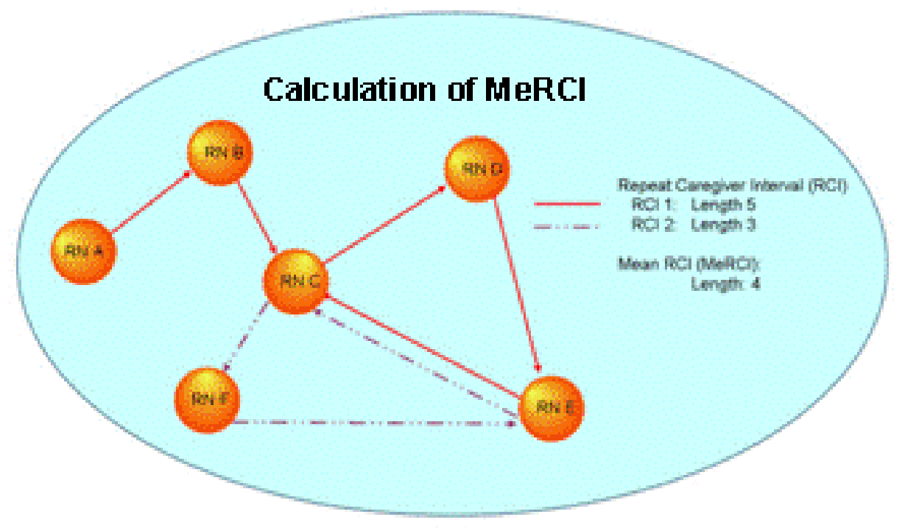
Methods

Using EHR data, we construct individual patient networks representing handoffs (edges) that occurred between nurses (nodes) who cared for the patient.



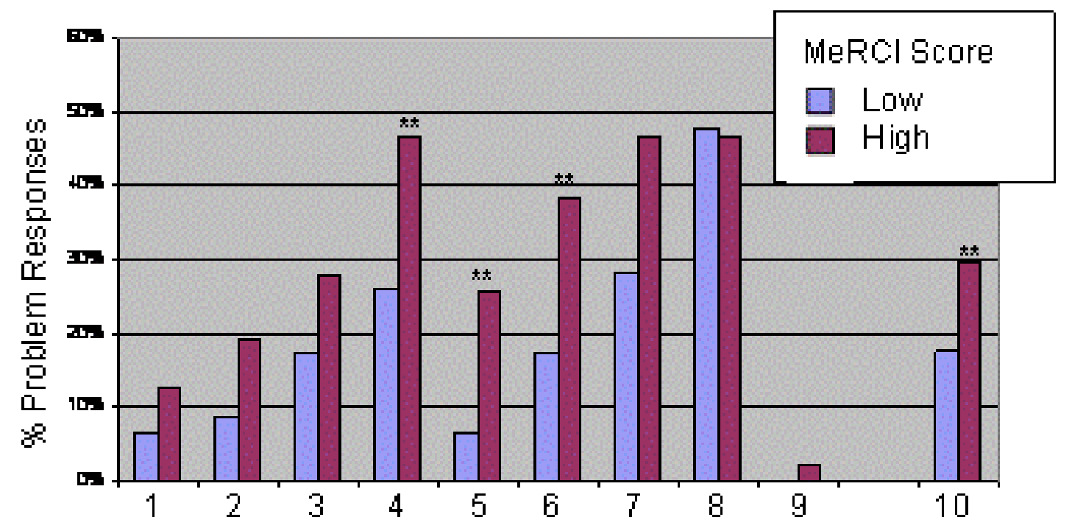
We calculated various network standard network statistics. We also developed our own measure of continuity of care, the mean repeat caregiver interval (MeRCI). From the first caregiver, we count the number of shifts before a repeat nurse. This is a Repeat Caregiver Interval (RCI). From the end of the previous RCI, this process is repeated iteratively, and MeRCI is the average RCI of the patient.

Family satisfaction with care for each patient was based on parents' responses to the Picker Institute's NICU Family Satisfaction Survey.



Results

High MeRCI scores were significantly correlated with poor scores for family satisfaction with care.



1 - RN's available to answer questions/concerns
2 - Understood RN's answers to questions
3 - RN's paid attention to your needs
4 - Had confidence and trust in RN's
5 - Knew which RN was in charge of baby
6 - If primary knew RN caring for baby
7 - RN on duty had enough info about baby
8 - Comfortable with RN taking care of baby
9 - Rated availability of RN's
10 - Overall RN problem score

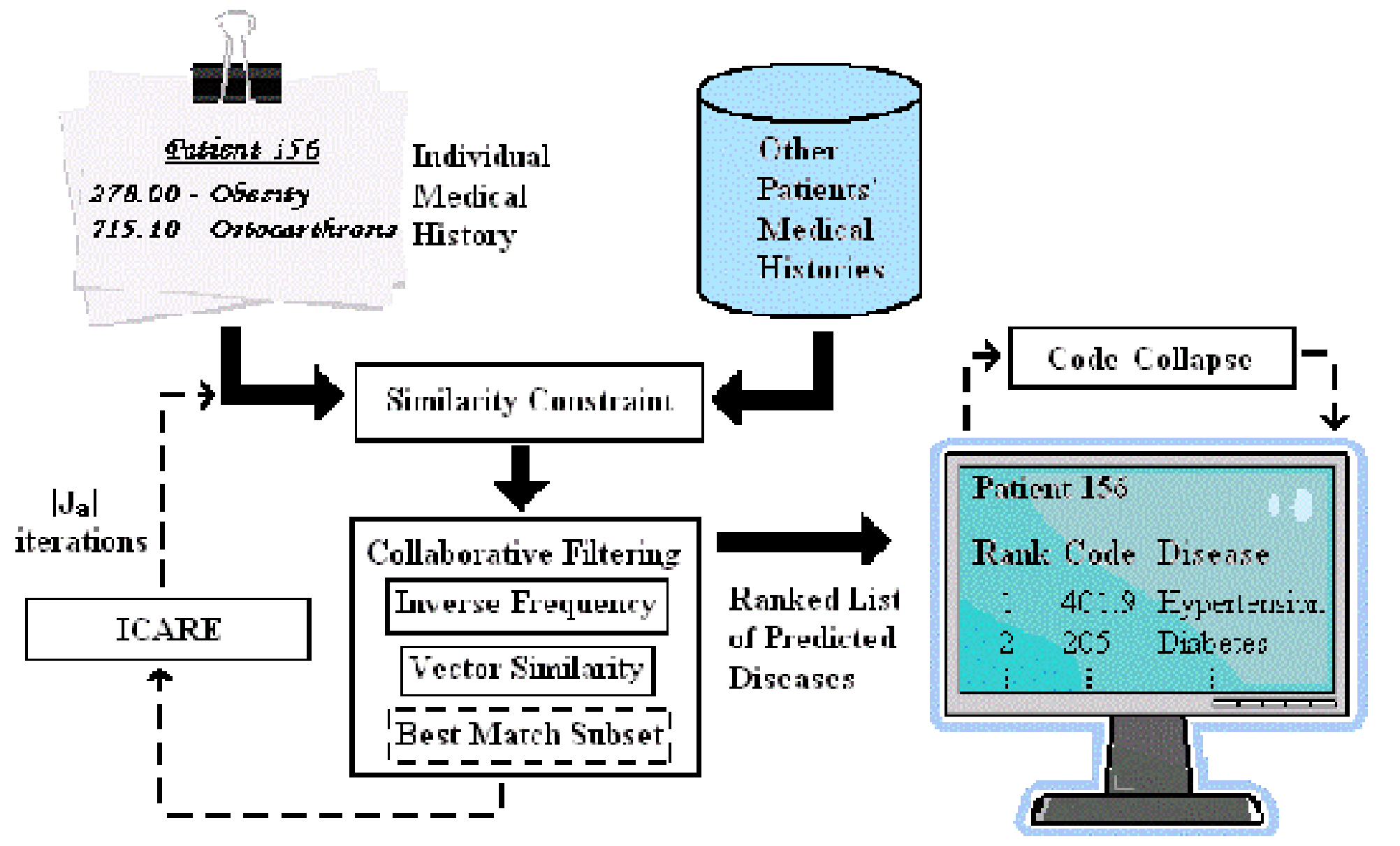
Collaborators: James Gray, DeWayne Pursley, Jane Smallcomb , Alon Geva (Beth Israel Deaconess Medical Center

The CARE Recommendation System for Personalized Healthcare

Many medical conditions have recognizable indicators before onset or preventable risk factors. However, universal testing is neither time nor cost efficient, and could cause medical harm. Currently, physicians use family and health history and physical examination to approximate the risk of a patient, guiding laboratory tests to further assess the patient's stage of health. However, these sporadic and qualitative "risk assessments" focus on only a few diseases and are limited by a particular doctor's experience, memory and time. We describe CARE, a Collaborative Assessment and Recommendation Engine, which uses patient medical history to generate a personalized risk profile for a patient. CARE is a comprehensive recommendation system that considers the experience of millions of patients and provides a personal answer to the question:

What are *my* disease risks?

The CARE Framework



The central component of CARE is collaborative filtering. In the same way that services like Netflix predict items that their users will enjoy, CARE predicts diseases that patients will develop based on the diagnoses of patients who had a similar medical history. It is very likely that other patients among millions have experienced genetic and environmental risk factors that closely mirror our own.

Data and Methods

Our data comprises of Medicare records of more than 13 million elderly patients. Each data record consists of a hospital visit, represented by up to ten ICD-9-CM diagnosis codes. We also have outpatient data for 800,000 patients within a large regional health system, also in ICD-9-CM format, spanning all age ranges.

We use collaborative filtering with vector similarity weighting. We also incorporate inverse frequency, meaning rare diseases influence similarity more strongly. Our Iterative version, ICARE, uses ensembles of multiple collaborative filtering rounds to isolate significant correlations and control common diseases. We also determine patient similarity based on the subset of consecutive visits from a training patients record that best matches the active patient (for whom predictions are being made). This allows the algorithm to use only the most relevant portion of each medical record, which also reduces noise and complexity. For each patient, the system outputs a ranked list of diseases from the highest risk score to the lowest.

Evaluation

Our strongest evaluation metric is coverage, the percentage of a patient's actual future diseases which a prediction is made and ranked. For evaluation, we usually limit the prediction list to the top 20 highest risk scores, a practical size for consideration by a medical professional.

Results

Our best method, the ensemble-based ICARE, captures 41% and 45% of all future diagnoses in the top 20 ranks for the Medicare and outpatient data, respectively.

Collaborators: Albert-Laszlo Barabasi, Nicholas Blumm (Northeastern University), Nicholas Christakis (Harvard University)

UNIVERSITY OF NOTRE DAME