IDS 560 ANALYTICS STRATEGY AND PRACTICE

THE BENEFIT IS CLEAR

4C HEALTH SOLUTIONS
MEDICAL EPISODE DETECTION

Team Members:
Komal Tejwani
Nancy Valdez
Pooja Narayanan
Santosh Kumar Ponnada

Healthcare Industry

Within the healthcare industry, claims are captured from numerous sources which include pharmacy, medical, facility claims, and dental administrators. Currently, the industry does not have standardized health care collection. Health records often are in different formats and on different systems. Smart healthcare is what many are referring to as a modern shape of healthcare. Ideally smart healthcare would include the following:

- Clinicians use technology to more accurately diagnose and deliver care
- Patient data is in one place and effectively accessible
- Cost-effective delivery models would bring healthcare to people who do not have it
- Appropriate treatments are delivered at a timely manner

The healthcare industry faces the current problem of how to advance processes, policies and capabilities to deliver savvy healthcare to everyone.

Introduction

4C is a healthcare solutions company that focuses on tackling the epidemic of healthcare fraud and waste for self-insured employers and government programs. 4C works with employers and administrators to identify and prevent healthcare payments fraud. They are using proactive analytics which includes machine learning and artificial intelligence to bring accountability and transparency for companies to monitor their health benefits. Due to many employers who use third-party administrators not having visibility on how their benefits programs are managed. 4C is hoping that with proactive analytics, they can give employers clarity in terms of their benefits programs. Using predictive analytics, 4C can identify questionable claims which result in fraud prevention. This happens in real time which allows the employer to ask for more information on the questionable claim instead of paying it right away.

Overview

The main motivation of this project was to construct complex communities' structures by employing graph theory and network analysis to detect clinical episodes of care for patients. In future, in the event that 4C chooses to go to a clinical subject matter to guide these episodes, they can utilize the proposed community discovery algorithms to recognize these episodes of care.

About the Data

In healthcare data sources, data standardization is a key pillar for efficient and meaningful use of the information and collaboration of healthcare professionals, care providers, insurers, and government agencies. Our medical claim data comprises details of the services provided, condition of patient, service date of claims, procedure, changes in procedure (if any).

The most important variable for our problem statement is the 'Current Procedural Technology' (CPT). CPT is an alphanumeric medical code set that is used to report medical, surgical, and diagnostic procedures and services to entities such as physicians, health insurance companies and accreditation organizations. Each claim is associated with a specific 'Payer Claim Control Number' - a number assigned by the payer to identify a claim, 'Employer ID' - the 4C customer who submitted the claim and the 'Plan administrator ID' – the unique identification ID of the payer.

Building the communities

Network theory and community detection methods were applied to medical claims data to visualize, analyze, and understand the medical episodes to make better data driven decisions based. A network graph is the graphical representation either symmetric relations or asymmetric relations between the nodes. Community detection in networks aids to identify the communities of procedures codes based on their common characteristics and similar behavior.

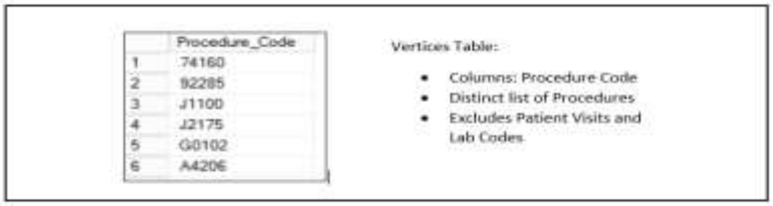
In our clinical claims data, the vertices and edges were defined as below:
Vertices - Distinct procedure codes (CPT)
Edges - All combinations of procedure codes (CPT)
Weights - Count of number of patients with same set of procedure codes.

*'Walktrap Community Detection' and 'Cluster Label Propagation Community' algorithm*s were used to identify communities and compare the authenticity of the results with each other.

The medical claims data was transformed by a bunch of SQL modules which made it compatible to run community detection algorithms on. First network graph and community detection were applied to the most frequent diagnosis code in our data - Type II Diabetes. Using the medical claims data, nodes and edges tables were created.
Given below is the stepwise explanation on how these files were created from the claims data:

- Filtering the data for an employer id and plan administrator id
- Filtering the claims for patients who were diagnosed with, for example 'diabetes'
- Removing procedure codes related to regular office visits and lab codes like drawing of blood samples
- Creating the edges table by finding all the combinations of procedure codes for all the patients
- Assigning the weight to each combination of procedure codes. The weight was given on the count of number of patients with the same set of procedure codes
- Creating the nodes table by finding the distinct list of procedures for all the patients in step 4

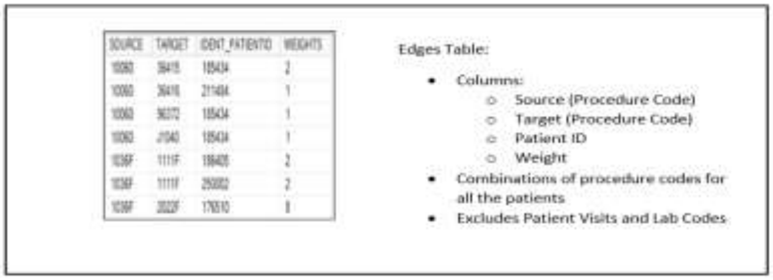
Below is a sample output of the data extraction process:



	Procedure_Code
1	74160
2	92285
3	J1100
4	J2175
5	G0102
6	A4206

Vertices Table:

- Columns: Procedure Code
- Distinct list of Procedures
- Excludes Patient Visits and Lab Codes



SOURCE	TARGET	IDENT_PATIENTID	WEIGHTS
10060	36415	185434	2
10060	36416	211484	1
10060	96372	185434	1
10060	J1040	185434	1
1036F	1111F	198405	2
1036F	1111F	250002	2
1036F	2022F	170510	8

Edges Table:

- Columns:
  - Source (Procedure Code)
  - Target (Procedure Code)
  - Patient ID
  - Weight
- Combinations of procedure codes for all the patients
- Excludes Patient Visits and Lab Codes

The communities obtained were cluttered and difficult to interpret, hence to reduce the number of nodes - *elbow method* was used. Elbow graph or inflection point is a method of interpretation and validation of consistency within cluster analysis. In this method, each node is plotted against the degree of node (number of edges incident to the vertex, with loops counted twice) and the elbow point is decided. The nodes below the above the elbow point (inflection point) were discarded from the data and community detection was again applied. The results are explained below.

Community Summary for patients diagnosed with diabetes:

- Total number of communities detected = 12
- Number of communities that were interlinked = 7
- Elbow point = 100 degrees

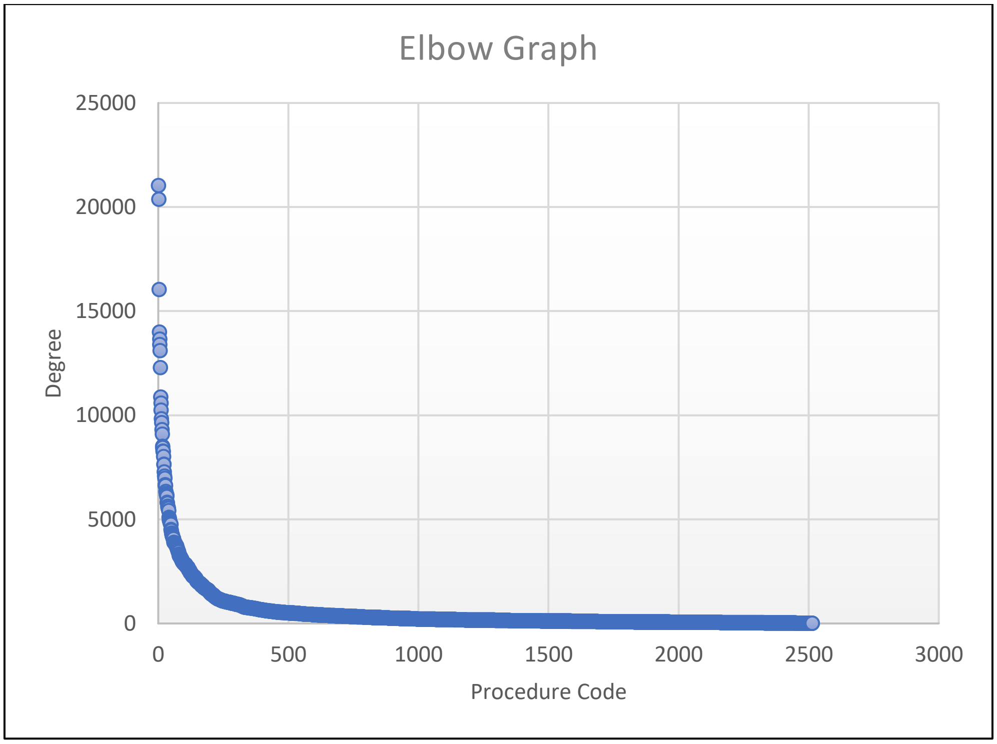


Fig 1: Elbow graph for patients diagnosed with Diabetes

From *fig 2*, we see several communities that were not interlinked and survived as individual entities. These were essentially considered as the noise in the dataset. The noisy procedure codes corresponded to codes for 'skin sealants, protectants, moisturizers, ointments', 'replacement battery for external infusion pump owned by patient', 'colonoscopy' etc. that were not related to the diabetic patients.

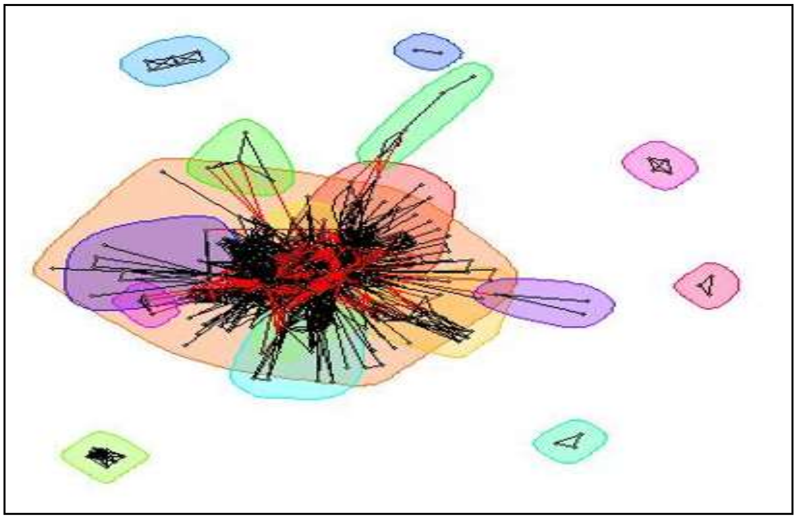


Fig 2: Network graph using Walktrap algorithm for patients diagnosed with Diabetes

We observed, the following key insights that could be taken from the above community graph:

Foot related community: Diabetes can cause some serious complications to feet. It is common for people with excess levels of glucose in their bloodstream for a long time can cause nerve damage that can cause tingling, pain and swelling of the feet. We observed, procedure codes relating to 'shoe molded to patient model' and 'shoe or custom-molded shoe' from which we gathered that diabetic patients had to wear custom made shoes.

Community with vision problem: Blurry vision is often one of the first warning signs of diabetes. Diabetic eye disease is a condition that can affect people suffering with diabetes. All forms of diabetic eye disease have the potential to cause severe vision loss and blindness. In our analysis, we found procedure codes corresponding to 'frames, purchases', 'polarization, any lens material, per lens' which helped us understand blurry vision could be sign of diabetes.

Community related to injection procedures: Diabetes generally arises when insulin is not produced on enough quantities or when the human body does not use the insulin well. There are several injections in the market which help produce insulin. A community with injection procedures like 'therapeutic, prophylactic or diagnostic injections, subcutaneous or intramuscular' used for diabetic treatment were also detected.

We focused on developing a network graph analysis approach in detecting a small community like diabetes. Later, we applied the algorithm to a larger dataset. The following table shows the top 10 diagnosis codes used for community detection:

Diagnosis Code	Description
25000	Diabetes
4019	Hypertension
2724	Hyperlipidemia (Condition in which there are high levels of fat particles in the blood.)
78650	Chest Pain
V7612	Malignant neoplasm of breast

4011	Benign essential hypertension
7242	Lumbago (Backache of the lumbar region or lower back, which can be caused by muscle strain or a slipped disk)
78900	Abdominal pain
5856	End stage renal disease
78079	Malaise and fatigue

Table 1: Description of Top 10 diagnosis codes in claims data

Community Summary for top 10 diagnosis codes:

- 16000 vertices (or procedure codes)
- Number of interlinked communities = 18
- Size of largest community is 1466
- Elbow point= 2000 degrees

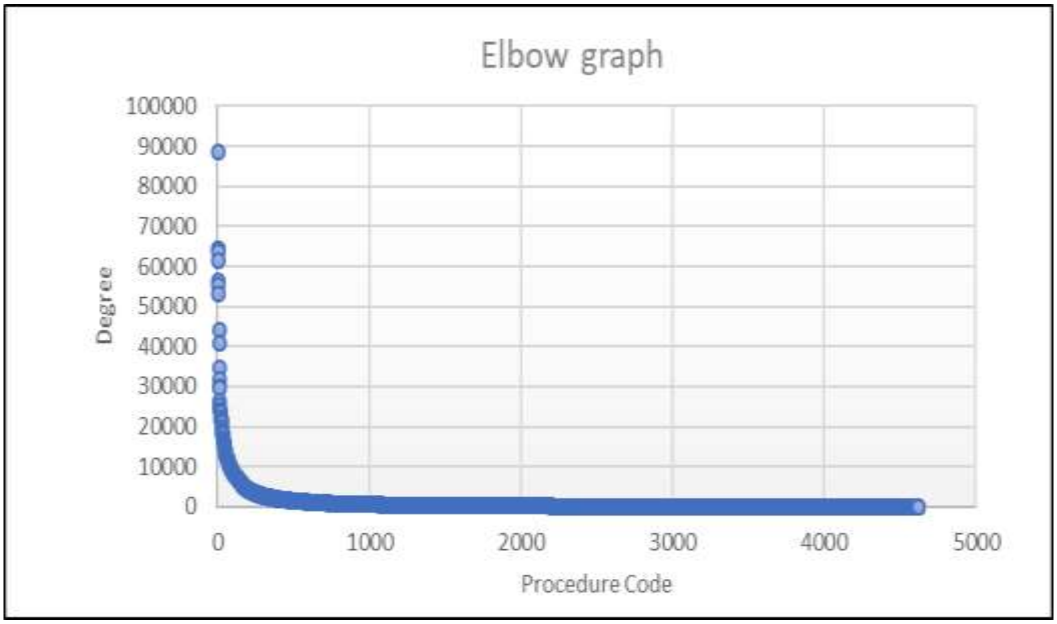


Fig 3: Elbow graph for patients diagnosed with top 10 diagnosis codes

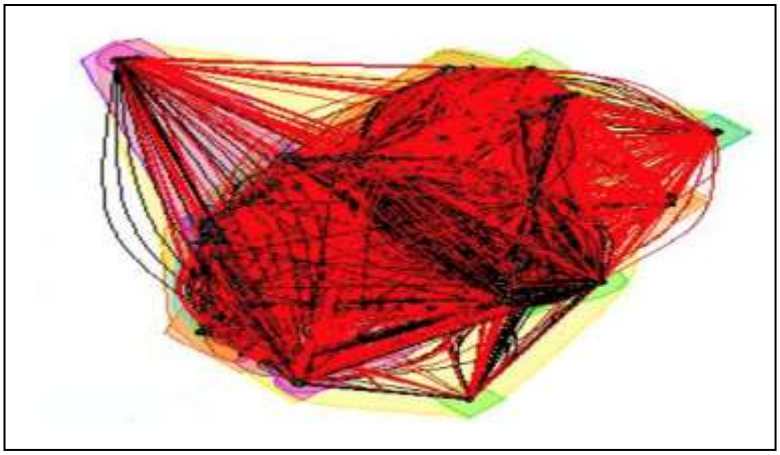


Fig 4: Network graph using Walktrap algorithm for patients with top 10 diagnosis codes

We observed, the following key insights that could be taken from the above community graph:

- One giant community of 1466 procedures codes were formed. This comprised of nursing, discharge, new patient preventive medicine services, manipulative treatment procedures, etc.
- The algorithm did not always give pure communities. For example, the second largest community comprised of heart problems and pediatric services. Similarly, skin related issues were grouped with ear infection.

The table below gives the description of a various communities:

#	Community Size	Description of Community
1	1466	General Services (Nursing + Discharge etc.)
2	35	Heart related and Pediatric services
3	19	Prosthetic Services
4	18	Integumentary (Skin) Tissues + Ear Infection
5	13	Nuclear Medicine Procedures
6	11	Cardiac Assist Procedures + Procedures for Hands and Fingers

Table 2: Description of top 6 communities based on size

After developing the communities, the patients were connected back to communities to which they belong. Since a patient can be a part of more than one community, this helped us understand how a patient developed different diseases through time. In addition, focusing on a single community helped us understand the most common procedures that doctors were following to treat a diagnosis.

Below is a summary of how patients are mapped to a community:

- Number of patients = 490
- Number of patients belonging to 3 different communities = 1
- Number of patients belonging to 2 different communities = 42
- Number of patients belonging to a single community = 447
- Number of patients belonging only to the ‘General Services’ community = 443

Future Scope

The future teams can try to work on the community detection approach to understand the following:

Risk Adjustment: Find communities of patients with chronic diseases that pose a higher risk to the insurance providers.

Cost Containment: Find a relationship and a reason as to why different doctors charge different bills (if any) for rendering the same type of service or disease. Finding these relationships will be beneficial for the insurance providers for them to adjust their cost.

Fraud Detection: Find relationships between patients and services forming complex communities' structures that could lead to potential fraud discoveries. In this project, we studied different types of relationships and focused on small but exclusive relationships and developed algorithms to detect these small and exclusive communities. These algorithms can be applied to a larger dataset and are highly scalable.

Appendix

R Code:

```
library("RODBC")

#Connect to DB

dbhandle <- odbcDriverConnect('driver={SQL Server}; server=CC-SQL16-IDS506; database=4C; rows_at_time = 1; trusted_connection=true')

#Load table has only diagnosis 25000 with lab codes and office visit removed

vSQL25<-paste(" SELECT * FROM VERTICES_25000",sep="")

eSQL25<-paste("SELECT * FROM EDGES_25000",sep="")

vDF25<-sqlQuery(dbhandle,vSQL25)

eDF25<-sqlQuery(dbhandle,eSQL25)


#table has 8 diagnosis code with office visits filtered and lab test filtered

vSQLAll<-paste(" SELECT * FROM VERTICES_COMBINED_TOP8_PRUNE",sep="")

eSQLAll<-paste("SELECT * FROM EDGES_COMBINED_TOP8_PRUNE",sep="")

vAll<-sqlQuery(dbhandle,vSQLAll)

eAll<-sqlQuery(dbhandle,eSQLAll)

close(dbhandle)


#Load Package iGraph & Package Network Graph

library("igraph")
library("network")

# For diagnosis 25000 - Turning networks into igraph objects

g25 <- graph.data.frame(d=eDF25,  directed = FALSE,  vertices=vDF25)


#Walktrap

wc <- walktrap.community(g25, weights=V(g25)$Weight, membership = TRUE)

plot(wc, g25,  vertex.size=2,  vertex.label= NA)

modularity(wc)


#Community detection based on based on propagating labels

clp <- cluster_label_prop(g25)

plot(clp,  g25,  vertex.size=2,  vertex.label= NA,  vertex.color="orange",  edge.arrow.size=.2,  edge.curved=0)

modularity(clp)


#For table with 8 diagnosis code with office visits filtered and lab test filtered

gAll <- graph.data.frame(d=eAll,  directed = FALSE,  vertices=vAll)


#Walktrap

wcALL <- walktrap.community(gAll,weights=V(gAll)$Weight,membership = TRUE)

plot(wcALL, gAll,  vertex.size=2,  vertex.label= NA)

modularity(wcALL)


#Community detection based on based on propagating labels

clpAll <- cluster_label_prop(gAll)

plot(clpAll, gAll, vertex.size=2, vertex.label= NA,  vertex.color="orange", edge.arrow.size=.2,edge.curved=0)

modularity(clpAll)
```

Sponsor Presentation:

https://prezi.com/view/zI5qVe52ZFVtPc3S61YJ/