# **Using NEO4j to find Influential Doctors**

By Steve Petrie

10/25/2019

**Table of Contents**

[**Using NEO4j to find Influential Doctors** 1](#_Toc22891796)

[**Overview** 3](#_Toc22891797)

[**Influence Theory** 3](#_Toc22891798)

[**Business Case** 7](#_Toc22891799)

[**Solution** 7](#_Toc22891800)

[**Detailed Process** 7](#_Toc22891801)

[**Results** 8](#_Toc22891802)

[**Demo Web site** 11](#_Toc22891803)

[**References** 11](#_Toc22891804)

**Table of Figures**

[Figure 1 - Node Centrality 4](#_Toc22891925)

[Figure 2 - Node Communities 5](#_Toc22891926)

[Figure 3 - Large Network with Communities 6](#_Toc22891927)

[Figure 4 - Doctor Network Diagram 9](#_Toc22891928)

[Figure 5 - 3D Network Diagram 10](#_Toc22891929)

## **Overview**

Many healthcare marketing strategies require knowing who the most influential doctors in a specialty or area, then recruit these doctors to promote your product. Historically, the process involved reviewing a doctor's academic credentials, a doctor who had written a lot of academic papers or spoke at numerous conventions was deemed an influential person or a Key Opinion Leader (KOL). This person was then recruited to promote the product by speaking about it at events and conferences. In many ways, the KOL process is just another mass media channel were information is disseminated to a group of doctors, in a broadcast fashion. This process can be quite subjective based upon the reviewer’s opinion of who is important. This traditional approach doesn't account for all aspects in which influence can be exerted.

The purpose of this paper is to define a Neo4j based process that relies heavily on peer-to-peer influence over the traditional KOL influence. Peer-to-peer influence is the influence a colleague has on a person's decision.

According to Dr. E.M. Rogers:

*Mass media channels are more effective in creating knowledge of innovations, whereas interpersonal channels are more effective in forming and changing attitudes toward a new idea, and thus in influencing the decision to adopt or reject a new idea. Most individuals evaluate an innovation not on the basis of scientific research by experts but through the subjective evaluations of near peers who have adopted the innovation. These near peers thus serve as role models whose innovation behavior tends to be imitated by others in their system* (Rogers , p. 6).

Healthcare marketing traditionally has treated doctors as individual entities, but doctors as described by Dr. Rogers, don't make decisions as important as prescribing a new drug in isolation, they are heavily influenced by the opinions of their peers who have already adopted the drug.

## **Influence Theory**

According to Dr. Bergman, *“Word of mouth is the primary factor behind 20 percent to 50 percent of all purchasing decisions,”* (Bergman, J., p. 7).

According to Dr. Hill,

*…consumers linked to a prior customer—adopt the service (or product) at a rate 3–5 times greater than baseline groups selected by the best practices of the firm’s marketing team. In addition, analyzing the network allows the firm to acquire new customers who otherwise would have fallen through the cracks, because they would not have been identified based on traditional attributes. (Hill,2006)*

According to Dr. G Weiman,

*Social influence theory within social communication theory was posited by Paul Lazarsfeld and colleagues in the 1940s and 1950s. Their focus was on the power of informal communication as a complement to the influence of mass media. They discovered that informal communication is widespread, and that certain people were more central and influential than others in a group. They termed these individuals ‘opinion leaders’, thereby instigating a major topic of research that confirmed, expanded, and refined this idea (*Weimann ,1994*).*

**Centrality**

Network Science and Neo4j can be used to find the these 'opinion leaders' by locating the most central nodes (doctors) in the network. It is easier for the nodes in the center to communicate to the rest of the network than nodes on the perimeter. For example, in the network below, the blue nodes are more centrally located than the red nodes, thus if the blue nodes accept an innovation, this innovation will spread faster through the network than if a red node accepts the innovation.

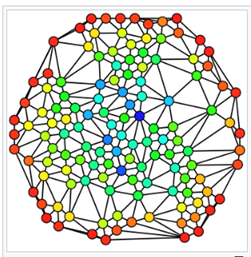


Figure - Node Centrality

**Community**

A second important influence metric is to determine the doctor’s community of peers, this is how the nodes naturally group together based upon the communication patterns. The groups below have stronger communications inside the groups than outside the groups.

It’s not unusual for groups to behave is a similar manner. Through statistical methods, this group coherence can be measured. We have found some groups are highly coherent adopting similar treatments whereas other groups have almost a random approach to treatments. A strong coherence may be reflective of strong communication patterns or an underlying treatment policy such as a clinic adopting a policy to us a specific drug. From a marketing approach finding the leaders of these highly coherent groups will greatly improve the effectiveness of the sales team.

Just as individuals exert influence each other, groups also exhibit influence over other groups. For example, if a respected clinic or hospital adopts a new treatment this will influence other groups in the area on the efficacy of the treatment. Thus, finding these influential groups is key to improving market sales.

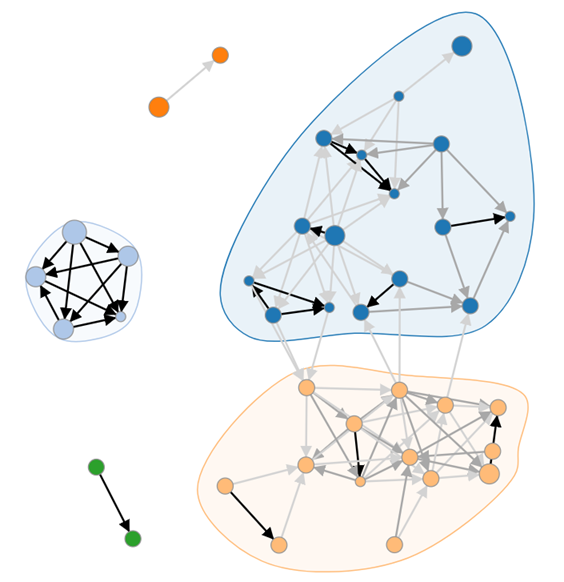


Figure - Node Communities

Below is a larger network of doctors. As you can tell from the diagram, it's not unusual for certain groups to control the communication pathways in the network and therefore influence how drugs are adopted. Focusing your marketing efforts on the blue clusters in this network will greatly improve sales.

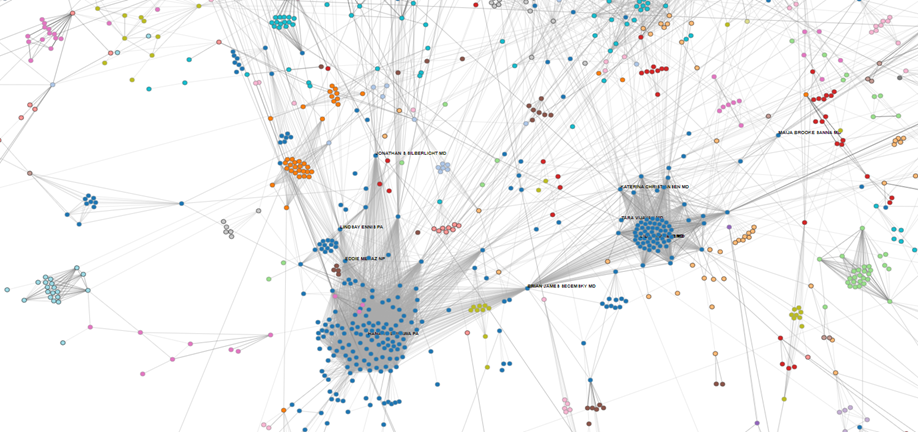


Figure - Large Network with Communities

## **Business Case**

A client was bringing a new drug to market but had a limited marketing budget. We suggested finding the most influential doctors in an area and market to them. They in turn via word of mouth would influence their peers on the efficacy of the drug and thus market the drug for the client.

The client was skeptical that we could measure doctor influence at all, let alone find the most influential individuals. After some negotiation, the client agreed to the contract on the condition we statistically prove that influence occurred. Therefore, the deliverables were: 1) target list of doctors rank by influence, 2) naturally occurring doctor groups, 5) visual diagram of the network in both 2D and 3D.

## **Solution**

Our approach to the problem was to map the relationships between all the doctors in a target area using NEO4j and then find the centrality and communities in this network. The most central of the doctors were the most influential. This is a common technique used in social science and network science.

## **Detailed Process**

1. **Obtain doctor data**, specifically data that contains relationships between doctors.
   1. Identify target doctors. This is a subset of all doctors that are specific to the project. It may be a specific specialty, a location, disease or drug.
   2. National Provider Identifier (NPI) data - demographic info from NPI registry. This also contains non-directional relationship data such as:
      1. Where the doctor works, two people working at the same clinic are assumed to know each other. This is determined by:
         1. Same formal name of the organization
         2. Same address
         3. Same phone number
         4. Distance between addresses (doctors in the same block probably know each other)
   3. Referral data – Centers for Medicare and Medicaid Services (CMS) lists doctor referral information, basically who is referring patients to whom. This is directional data that also contains the number of patients referred. This number could be used to define the strength of the relationship plus the direction.
   4. Co-Author data from medical journals - This data can be obtained from Pubmed data.
   5. Open payments data - This can be used to determine who the pharma industry considers influencers by the amount of money spent on the doctors.
   6. Insurance Claims data - This data lists which prescription was written by which provider on which day. This could be one or more drugs, ideally similar to our client's market. This information proved key in proving whether influence occurred or not.
2. **Generate Doctor Relationships** - this is a network that maps all the possible communications paths between doctors using:
   1. CMS \ NPI relationships
   2. Academic paper co-author relationships
3. **Import target doctors into a NEO4j graph database**
4. **Import doctor relationships into a NEO4j graph database**
5. **Run doctor network centrality calculations on neo4j**  - write results to neo4j doctor nodes
   1. Betweenness
   2. Closeness
6. **Calculate statistical significance of influence** – After the network is mapped, we look to see how information spreads through the network. Using the insurance claims information, we answered the question, “Are doctors more likely to prescribe a medication if they know someone who has prescribed it?” Using the claims data as a time series, for each new prescription we determined whether the doctor was linked to a peer who has already prescribed the medication or whether this was a random event.
7. **Generate visuals using D3**

## **Results**

Below are the results of using the above process in the targeting psychiatrists working in Philadelphia.

**Statistical Significance**

Using the process described in item 8 above we determined that there was that a doctor was more likely to prescribe a medication if their peer had already prescribed the same medication. His was statistically significant.

**Top 10 target list**

Table - Doctor Target List

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **NPI** | **Last Name** | **First Name** | **Address\_1** | **Address\_2** | **Rank** | **Scripts** | **Payments** |
| 1215149661 | SARKAR | PRIYANKAR | 5501 OLD YORK RD | OFFICE OF ACADEMIC AFFAIRS | 1 | 2146 | $75.86 |
| 1437313186 | SANUCK | NEIL | 100 E LEHIGH AVE | MAB BLDG, SUITE 105 | 2 | 0 | $0.00 |
| 1962441725 | OTERO-PEREZ | GUILLERMO | 101 E OLNEY AVE | SUITE 400 | 3 | 2403 | $0.00 |
| 1467426247 | KHAN | MAHRUKH | 5501 OLD YORK RD |  | 4 | 418 | $95.54 |
| 1053571000 | MARGERY | SOLANGE | 833 CHESTNUT ST | SUITE 210 | 5 | 0 | $0.00 |
| 1184620130 | IGE | ABAYOMI | 1315 WINDRIM AVE |  | 6 | 2453 | $319.44 |
| 1528099355 | RICHARDSON | THOMAS | 101 E OLNEY AVE | SUITE 400 | 7 | 198 | $502.02 |
| 1992909733 | CRUZ LUNA | NORMA | 5501 OLD YORK RD |  | 8 | 287 | $0.00 |
| 1609900455 | MAGNANI | GRETCHEN | 4641 ROOSEVELT BLVD | FRIENDS CAMPUS SCATTERGOOD BLDG | 9 | 0 | $0.00 |
| 1154455798 | ALHOMSI | MOTAZ | 27 E MOUNT AIRY AVE |  | 10 | 3559 | $564.20 |

**Network diagram**

Below is a two-dimensional diagram of the Philadelphia doctor network. You can clearly see the natural groupings of the doctors. The size of the nodes is the number of prescriptions the doctor has written. The bold names are the most influential doctors I this area. The link weight is strength of the relationship between the doctor. In the diagram below, Dr. Sarkar is the most influential doctor. He is in the center of the large blue group. It is interesting to note, that while he doesn’t write the most prescriptions, he is connected to doctors to who do. Thus, convincing the core group of doctors in the center of the group to use a specific product has a high probability of that product being adopted by the whole group.

The smaller light blue group to the right is highly coherent with lots of communications between the doctors. I would suspect this is a hospital or clinic. While they don’t write as many prescriptions as the dark blue clinic, I would suspect information to pass quickly in this group. Thus one or two doctors adopting a new drug has the potential of convincing a lot of doctors on its efficacy.

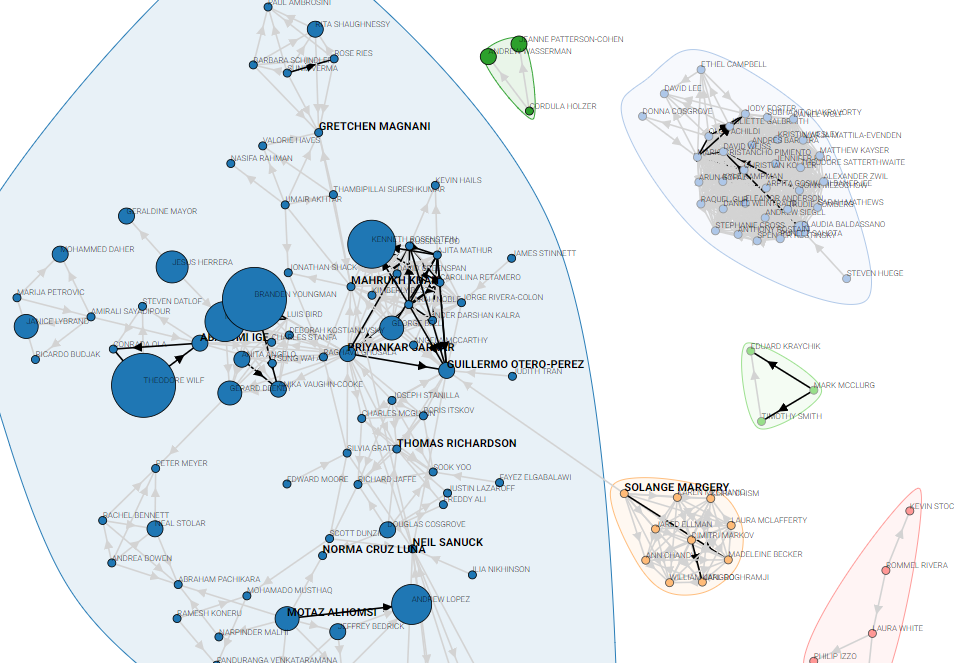


Figure - Doctor Network Diagram

**3D Network diagram**

Below is a 3D version of the network diagram. The purpose any data visual is to provide some insight into the underlying data in order to make between business decisions. In most cases, you can gain more insight from a 2D network than a 3D network; however, when the network is highly complex with thousands of nodes and hundreds or thousands of links a 3D version does allow you to fly in and around the data find very specific data points and their localized structure. Thus, between the 2D and 3D diagrams you can gain significant insight into the data that you might not gain from looking at the raw data.

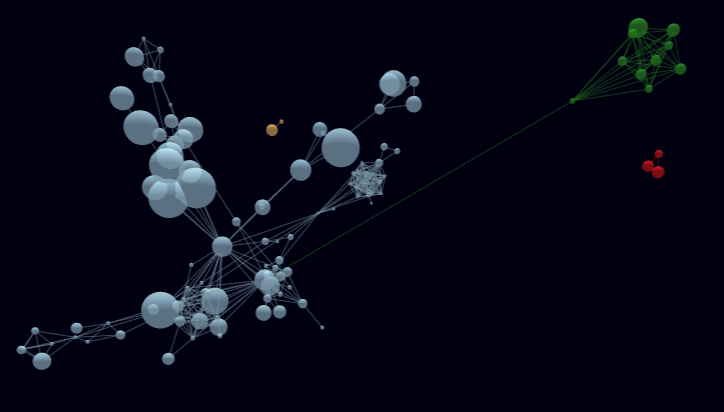


Figure - 3D Network Diagram

## 

## **Demo Web site**

The web site: <https://CGN38.github.io/bluejacket_network> contains an interactive version of the results table, 2D network and 3D network along with an online version of this paper.

## **References**

Bergman, J. (2013). Contagious: Why things catch on. New York: Simon & Schuster

Shawndra Hill, Foster Provost and Chris Volinsky, (2006). Network-Based Marketing: Identifying Likely Adopters via Consumer Networks, Statistical Science 2006, Vol. 21, No. 2, 256–276

Rogers, E. M. (2005). Diffusion of innovations. NY: Free Press.

Weimann, G. (1994). The influentials. Albany, NY: State University of New York Press.