Application of Social Network Analytics to Assessing Different Care Coordination Metrics

Ahmed F. Abdelzaher[1(✉)], Preetam Ghosh[2], Ahmad Al Musawi[3], and Ju Wang[1]

[1] Virginia State University, Petersburg, VA 23806, USA
{amohammed,jwang}@vsu.edu
[2] Virginia Commonwealth University, Richmond, VA 23084, USA
pghosh@vcu.edu
[3] Thi Qar University, Nasiriyah, Iraq
almusawiaf@utq.edu.iq
http://www.linkedin.com/in/ahmed-abdelzaher-322899119
http://www.linkedin.com/in/preetam-ghosh-5441502
http://www.linkedin.com/in/ahmad-al-musawi-577410141

Abstract. Social network analytic approaches have been previously proposed to identifying key metrics of physician care coordination. Optimizing care coordination is a primary national concern for which yields significant cuts in medical care costs. However, the proposed metric-termed 'care density' for estimating care coordination is not completely accurate. Our objective is to compare the accuracy of the previously proposed 'care density', with our proposed 'weighted care density', 'time varying care density', and 'time varying weighted care density' in terms of predicting the cost of care. Our proposed metrics are based on the former care density, however, takes other variables into consideration, mainly patient hospitalization time frames and number of physician visitations. Our findings suggest that physicians coordinating over short time spans spike the cost of care above normal.

Keywords: Social network analytics · 2 mode bipartite networks
Support vector regression · Incidence matrix

1 Introduction

Social networks belonging to a category of complex systems termed 'scale-free' [15], are known to follow a power law distribution for which the probability $p(K)$ for nodes to have neighbors is of the form $p(K) \sim K^{\gamma}$. The power law suggests that the evolution of such networks occurs in a sparse [7] manner, but more importantly, they exhibit "topological patterns". Similarly with other naturally occurring topologies, such as networks describing metabolic reactions in the animal cell, the World Wide Web and gene regulatory networks [4]. Essentially, such

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10914, pp. 151–160, 2018.
https://doi.org/10.1007/978-3-319-91485-5_11

network interactions exhibit particular behaviors having inner construction that distinguish there structural properties, and are not to be equaled with randomly generated networks using the famous Erdös-Rényi model [8] and the small world properties model of Watts and Strogatz [17].

The concepts of social network analytics have been applied extensively in the medical field attempting to estimate the level of collaboration among physicians sharing patients [14] and the factors affecting their influential views of primary medical practices [10]. Patient sharing increases the chances for interactive communications as well as higher levels of information exchange [5] and increase their chances of receiving efficient synchronized care. In fact, care coordination has been identified as one of the nation's most concerned area of research [1,3,13]. This motivated our work here, for which we try to understand the correlation between different metrics that can be useful in evaluating the level of care coordination in the medical field.

In doing so, we examine the correlation between the previously proposed care coordination measure, 'care density' [14], and the costs of medical care for patients admitted for suffering from pneumonia –a disease for which the estimated deaths combined with influenza was 53,582 for 2009, and for which care coordination is likely important [2]. We try to enhance the accuracy of the proposed metric by considering additional variables to the measure, mainly, the time frames for which the patients have been admitted and the number of distinct patient-physician visits. We propose, namely: (1) the weighted care density, (2) the time varying care density, and (3) the time varying weighted care density, in addition to the former care density and examine their cost of care correlations as well. Though previously suggested that increased synchronized care yields significant cuts in costs of care [14], our findings suggest that this might not be the case when considering hospitalization time frames of the patients.

In order to determine the most accurate care coordination measure, we extracted additional data per patient as discussed in the Data section, which are relevant for constructing support vector regression (SVR) models as discussed in the Methods section. Our SVRs model the effects of the different patient features (extracted data) when combined with the care coordination metrics to predicting the cost of care. We construct 4 different SVRs, each considering different care coordination metrics as distinct features per patient and the corresponding accuracy of the models are recorded in the Results section.

2 Data

Our records for patients suffering from pneumonia were provided by The Medical Center of Virginia (MCV). The data contains two spread sheets: (1) one contains 33920 records of different patient-physician visitations, (2) the other indexes the different codes for the type of patient discharge. The data encompasses 2324 pneumonia patients and 1506 providers operating between the dates: the 30th of September 2007 and the 11th of April 2008. For the time being, we employ patients feature extractions from the first document and we consider all types

of patient discharges the same. Computational programs have been constructed for parsing (and calculating if necessary) the different features pertaining to each patient, their notations and descriptions are portrayed in Table 1. Other information including race and gender of the patient, specialty of the provider, the patient's source of admission and the payer IDs are available within the spreadsheet but were not considered.

Table 1. Patient features

Data notation	Content description
#Doctors	Number of different physicians visited by patient
#Interventions	Number of different physician-patient visitations
#RX	Number of different medical prescription type visits
#LAB	Number of different lab orders type visits
#ADM	Number of other types of medical admitted visits
LOS	The length of hospitalization in days
Cost	The total cost of hospitalization
t	The time the patient admitted relative to t_0*

* t_0 is the time when patient was first admitted

3 Methods

3.1 2-Mode Bipartite Social Netwok

Calculating the care density requires forming a social network of patients and physicians known as the 2-mode bipartite social network- a network formed by assigning physicians as groups and patients they treat as subscribers to the groups [11,12,16]. Therefore, our network can be expressed by a set of doctors D and a set of patients P, together $(P \cup D) = N$, were N denotes the set of nodes of our network. We denote a binary relationship "Visits" as;

$$R_v = \{< p, d > | \exists\, p\, visits\, d\, or\, d\, provides\, for\, p\}, \tag{1}$$

wherein $p \in P$ and $d \in D$. Hence for $< i, j > \in R_v$, the value of the incidence matrix $G_{ij} = 1$ for a visit and 0 otherwise. In the weighted version of the 2-mode bipartite graph, the value of G_{ij} would be the number of times patient i visited doctor j. Note that G's number of rows equal to the number of patients and columns equal to the number of physicians.

3.2 Care Density

The Care Density (CD) [14] calculates an approximated value for the collaboration among doctors that a particular patient has visited during his entire stay

in the hospital. For a particular patient p who has visited n_d doctors, CD can be computed as follows;

$$CD_p = \frac{\sum_{i=1}^{m} w_{p,i}}{n_d(n_d - 1)/2} \tag{2}$$

CD has proven to be correlated to the reduction in the mean charges of hospitalization as reported in a study involving patients with Diabetes and Congestive Heart Diseases [14]. Consider Fig. 1, patient $p2$ visits $d1$, $d2$ and $d3$ during his/her entire stay which yields a 3 possible provider combination, therefore the denominator is 3. Between $d1$ and $d2$, 2 patients are shared, similarly with $d2$ and $d3$, while $d1$ and $d3$ share 1 patient. Therefore the physicians share a total of 5 patients over 3 possible pairs of doctors, CD = 1.67. We hypothesize that CD is not an accurate measure for physician's collaboration as it does not consider the different time windows at which patients were admitted and discharged.

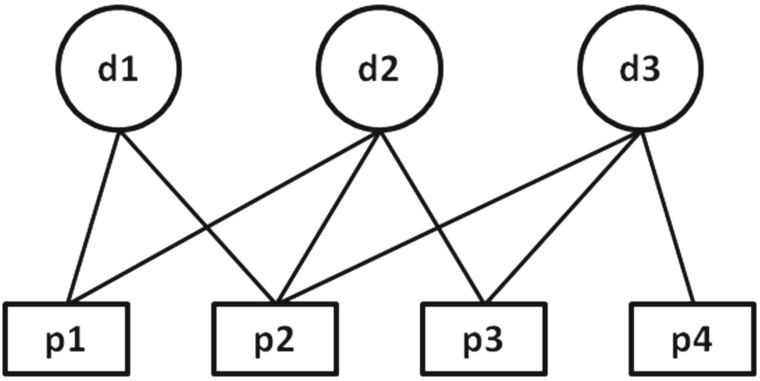

Fig. 1. A representation of a 2-mode bipartite graph between doctors (d1-d3) and patients (p1-p4).

3.3 Time Varying Care Density

Now, consider introducing two time windows $t_0 - t_1$ and $t_1 - t_2$ in Fig. 2, such that p1 and p2 have been admitted and discharged before t_1, and consider $p3$ and $p4$ being admitted after t_1. Doctors $d2$ and $d3$ share $p3$ after $p2$ was discharged and the effects of their collaboration post $p2$s treatment should not be incorporated in the CD of $p2$. Similarly with $p3$, CD should not account for the effects of the doctor's collaboration before $p3$ got admitted. Moreover, it makes no sense to try correlating the charges of a particular patient based on his provider's activity after a patient has been discharged and billed. Therefore the existing implementation of CD over estimates the collaboration of the doctors, which lead us to introducing the Time Varying Care Density (TCD). TCD excludes doctor's collaborations that occur outside the patient's time window. Essentially, $p2$ will lose the effect of doctors $d2$ and $d3$ sharing patient $p3$, which brings down the CD from 1.67 to 1.33, or TCD = 1.33.

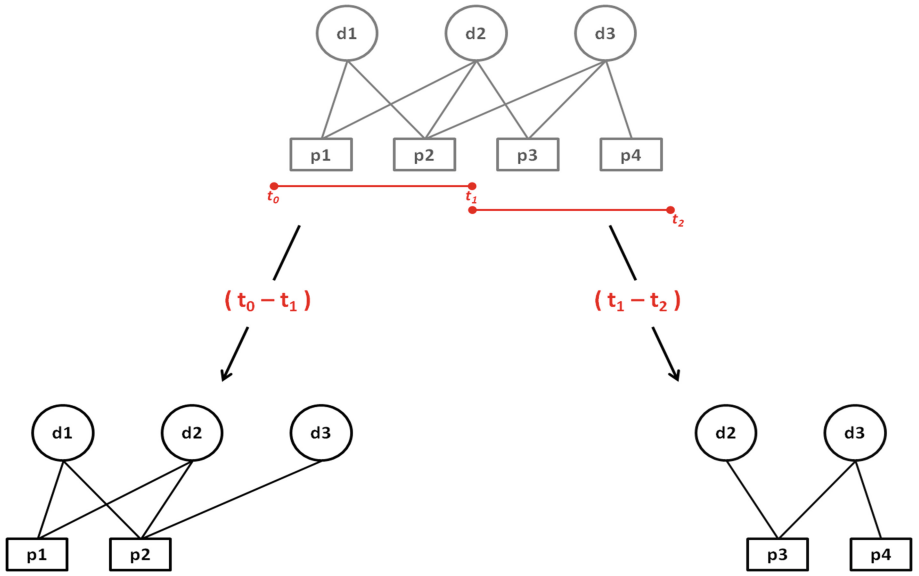

Fig. 2. A representation of the effects of introducing time windows to the social network of Fig. 1.

3.4 Weighted Care Densities

So far (in the above examples) we have considered the un-weighted incidence matrix, that is, we have yet to considered the number of visits patients have committed to doctors. In other words, we have considered CD to be the same for the following scenarios of 2 doctors sharing a patient: (Scenario 1) $p1$ visits $d1$ 9 times and $d2$ 1 time, and (Scenario 2) $p1$ visits $d1$ 5 times and $p1$ visits $d2$ 5 times. Both scenarios have a sum of visits equal to 10 but ideally $d1$ and $d2$ have a tendency to collaborate more in the second, i.e. CD should be higher in the first scenario.

To account for the weights (visits), we consider the collaboration between 2 physicians to be a maximum when both doctors have equal visits for all their common patients, and decreases as the difference between the numbers of visitations increase. The collaboration between a pair of doctors can be estimated using

$$w_{p,\,i} = 1 - 2\left|\frac{1}{2} - \frac{v_1}{v_1 + v_2}\right| \tag{3}$$

where $w_{p,\,i}$ is the value of a single operation of the summation of the numerator of CD_p. v_1 and v_2 constitute the sum of visits for the common patients between $d1$ and $d2$ respectively. From here on we refer to the weighted care density as WCD and the time varying weighted density as TWCD.

3.5 Support Vector Regression Modeling

SVR is a multidimensional modeling technique which tries to fit the best fit curve or hyper plane to recorded experimental data with minimal error difference between the actual data and the curve's estimation. Normally, the recorded data will depend on a set of features (or dimensions) such as in our case, were we try to model the total charges per patient and their dependence on the patient's features listed in Table 1. Note that time windows are not a patient feature used in the SVR, but are required for calculating the different care densities which are relevant patient dimensions.

As a useful method of modeling, many tools support SVRs. WEKA [9] is a user friendly GUI which requires the user to list the data and their corresponding features in a column format having the actual values (values to be estimated by the hyper plane) on the last column of the spreadsheet. Other tools come in the form of programming libraries such as LibSVM [6], which requires the user to write coding commands in addition to separating the actual values in a separate vector in the same order of the feature matrix. The feature matrix X should be of size $p \times f$ for p patients and f features such that X_{11} represents the value of the 1st feature of the 1st patient; X_{12} represent the value of the 2nd feature of the 1st patient and so up to X_{pf}. The actual data should be stored in the same order of the patients in a vector y of size $1 \times p$. Ultimately the SVR gives a best fit with approximated values;

$$y_a = wX + b \tag{4}$$

were w is a $1 \times f$ weights vector, which represents the relative influence of each feature on the SVR, and b is the bias which represents a constant shift to the curve. The above process is referred to as training, while testing w and b on foreign data is referred to as prediction.

In order to test the SVR, we use a 10 k-fold cross validation technique. The data is divided into 10 different chunks of 90 percents and 10 percents. The training is performed on the 90% portion and the corresponding w and b are used for prediction of the remaining 10%. The error of the 10% is accumulated each time using;

$$e = e + \left[\frac{||y - y_a||}{y} \times 100 \right] \tag{5}$$

then averaged across the 10 folds to give the cross validation error. We compare the 4 SVR accuracies in terms of predicting the cost of hospitalizations using the patient features listed in Table 1 (excluding cost and t) plus an additional feature for each model, that is the different care densities (CD, WCD, TCD, TWCD).

4 Results

Results of the un-weighted and weighted mean care densities, recorded in Table 2, show that higher care density receivers have lower average costs of hospitalizations. Because the care densities are skewed, we divide the statistical brackets

into tertiles of almost equal numbers of data points: lower, middle and upper. In both (CD and WCD) cases, there is no strong correlation between the average age of the patients and the average CD values.

Table 2. Features of pneumonia patients, stratified by care densities

	Un-weighted			Weighted		
Mean (SD)	Lower	Middle	Upper	Lower	Middle	Upper
N	768	765	751	739	755	790
Care density	3.32 (1.37)	7.53 (1.31)	20.22 (12.86)	2.68 (1.09)	6.03 (1.06)	13.94 (6.05)
Age	48.82 (23.26)	47.52 (24.50)	50.85 (21.93)	49.34 (22.89)	47.58 (24.52)	49.96 (22.5)
#Doctors	11.77 (9.08)	12.09 (7.94)	6.4 (4.57)	12.81 (10.97)	12.71 (8.30)	6.83 (4.83)
#inter-ventions	16.74 (13.87)	17.01 (12.39)	8.63 (7.02)	18.47 (17.18)	18.00 (12.97)	9.22 (7.41)
LOS	15.39 (15.03)	13.01 (12.4)	6.22 (6.79)	20.19 (44.14)	14.09 (13.88)	6.84 (7.68)
Charges*	102K (119K)	96K (113K)	39K (64K)	141K (252K)	107K (137K)	44K (72.8K)

*K: $\times 1000$

On the other hand, considering the time frames reverses every correlation mentioned above as depicted in Table 3. The mean costs of hospitalizations as well as mean number of interventions and mean LOS increase with respect to an increase in the TCD values. The higher the LOS, the higher the TCD simply based on the way TCD is calculated. Moreover, there is a direct correlation between the mean age and the mean TCDs, i.e. more care is required for elderly patients. It is also interesting to notice that the standard deviations of the CDs for each tile decrease as we add more variables to the metric, meaning $SD_{CD} > SD_{WCD} > SD_{TCD} > SD_{TWCD}$.

In order to determine the most suitable metric, we constructed an SVR using the 6 features: #Doctors, #Interventions, #RX, #LAB, #ADM, and LOS and compared it with the other 4 SVRs discussed above that have an additional CD as a 7th feature. As displayed in Table 4, All 4 SVRs show less cross validation error prediction than that of the SVR which excludes the CD. Though the difference in cross validation error is not so significant, we can still see a slight decrease in the cross validation errors as we consider more variables (time and weight) to affect the CD outcome.

Table 3. Features of pneumonia patients, stratified by time varying care densities

	Un-weighted			Weighted		
Mean (SD)	Lower	Middle	Upper	Lower	Middle	Upper
N	754	773	759	756	781	749
Care density	1.16 (0.3)	1.81 (0.21)	3.27 (1.07)	0.99 (0.26)	1.55 (0.18)	2.78 (0.9)
Age	46.7 (24.63)	48.3 (23.79)	52.1 (20.99)	47.31 (24.22)	48.47 (24.04)	51.42 (21.28)
#Doctors	8.64 (7.21)	9.45 (6.98)	12.25 (8.93)	8.65 (7.25)	9.67 (7.04)	12.05 (8.94)
#inter-ventions	11.78 (10.91)	13.13 (10.63)	17.59 (13.85)	11.96 (10.97)	13.43 (10.75)	17.17 (13.9)
LOS	8.91 (10.07)	10.57 (11.92)	15.26 (14.4)	8.86 (10.06)	10.84 (12.22)	15.1 (14.24)
Charges*	49.6K (75.3K)	67.5K (90.9K)	120KK (130K)	50.1K (75.7K)	68.8K (94K)	119K (128K)

*K: $\times 1000$

Table 4. Prediction errors for the different SVRs

SVR type	Cross validation error
Excluding CD	40.1%
Using CD	37.7%
Using WCD	37.7%
Using TCD	36.7%
Using TWCD	36.1%

5 Conclusion and Discussion

The original hypothesis by Pollack et al [14] is true when the care density values do not consider the time windows, however the time windows flaws the hypothesis. The original hypothesis considered an expanded time window that does not account for the stress exerted by the physicians, in fact, the effort exerted by a pair of doctors is always the same. However, collaborating in tight time spans reveals the urgency of the patient, and increases the stress amongst the physicians, which should normally bump up the cost of hospitalization.

The weighted time varying care density is the most accurate metric for assessing the physician collaboration. All 4 SVRs that consider one of the CDs as a 7th feature are more accurate than the 6 feature SVR, however, the SVR

which considers the TWCD is most accurate. Moreover, TWCD accounts for the visitations as well as the time windows, which give a more accurate estimation of the relative efforts exerted by physician pairs. The final costs of hospitalization should not be accountable for the activities of the physician past the patients discharge dates. Furthermore, TWCD gives a fair correlation between the age of the patient and the urgency of the disease.

It is important to note that the weight vector w of the SVRs show that the relative effect of the LOS feature on the SVR supersedes the rest of the features by a huge margin (including CD). This explains the slight, but not significant decrease in errors as we add more variables to the care density metric.

A more accurate SVR modeling approach would account for many other features which can influence the patient expenses. Firstly, the type of payer or insurance policy can be grouped to fall under 3 or 4 types of insurance policy, for which we can add a feature per type of policy, or model the data separately according to which policy type they fall onto. However for this kind of analysis we would require 3 to 4 times the amount of patients at hand. Similarly with the type of discharges, a study can be made to add a feature for discharges that are similar, such as; (1) discharged to another short term hospital and (2) transferred to another type of inpatient care institution. Thirdly, a feature can be added to the SVR which assesses the severity of the Pneumonia, for instance, a moderate case can take a value of 2.0 and a severe case a 5.0.

Acknowledgements. This work started as a class assignment, initially supported under the leadership of John A. Palesis, Ph.D, of Virginia Commonwealth University, and Johnathan P. Deshazo, Ph.D, of The Medical Center of Virginia. Without their support and guidance, the execution of this work would have been unlikely.

References

1. National Priorities and Goals: Aligning Our Efforts to Transform America's Healthcare. National Quality Forum, November 2008
2. 2011 national healthcare quality report (2011)
3. Adams, K., Corrigan, J.M.: Priority Areas for National Action: Transforming Health Care Quality. The National Academies Press, Washington (2003)
4. Albert, R., Jeong, H., Barabasi, A.: Error and attack tolerance of complex networks. Nature **406**(6794), 378–382 (2000)
5. Barnett, M., Landon, B., O'Malley, A., Keating, N., Christakis, N.: Mapping physician networks with self-reported and administrative data. Health Serv. Res. **46**(5), 1592–1609 (2011)
6. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27 (2011)
7. Del Genio, C.I., Gross, T., Bassler, K.E.: All scale-free networks are sparse. Phys. Rev. Lett. **107**, 178701 (2011)
8. Erdös, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. **7**, 17 (1960)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)

10. Keating, N., Ayanian, J., Cleary, P., Marsden, P.: Factors affecting influential discussions among physicians: a social network analysis of a primary care practice. J. Gen. Intern. Med. **22**(6), 794–798 (2007)
11. Luke, D., Harris, J.: Network analysis in public health: history, methods, and applications. Annu. Rev. Public Health **28**, 69–93 (2007)
12. Newman, M.E.J.: Networks: An Introduction. Oxford University Press, Oxford (2010)
13. U. D. of Health and H. Services. 2011 report to congress: National strategy for quality improvement in health care (2011)
14. Pollack, C., Weissman, G., Lemke, K., Hussey, P., Weiner, J.: Patient sharing among physicians and costs of care: a network analytic approach to care coordination using claims data. J. Gen. Intern. Med. **28**, 459–465 (2012)
15. Vázquez, A., Dobrin, R., Sergi, D., Eckmann, J.P., Oltvai, Z.N., Barabási, A.L.: The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc. Natl. Acad. Sci. U.S.A. **101**(52), 17940–17945 (2004)
16. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
17. Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)