Modeling Network Formation in an Online Health Community

Saumik Narayanan

Submitted under the supervision of Dr. Lana Yarosh to the University Honors Program at the University of Minnesota-Twin Cities in partial fulfillment of the requirements for the degree of Bachelor of Science, *magna cum laude* in Computer Science.

May 15th, 2019

Abstract

With the rise of the internet, online health communities have emerged as a resource for people with various medical conditions to form health-based connections. Like more traditional social media networks, many of these virtual connections are based on pre-existing connections formed outside the online health community. However, one major goal of these online health communities is to introduce users to others in similar situations, to provide and offer support, and to offer a platform for a meaningful, personal exchange of information (Neal 2006). This research aims to understand the reasons why new interactions are formed between users of online health communities. Our analysis of user interactions extends from previous research on network formation using conditional logit models, where interactions are modeled as individual choices made by users. Specifically, we aim to find and understand which features best predict the formation of a new connection between two users, with a focus on features derived from the graph itself, and features unique to the domain of online health communities. After creating several conditional logit models based on these features, we analyze the results to determine which features best predict the creation of an interaction. We find that the strongest predictor of a connection being formed is the degree of prior interactions with general visitors, while features from prior journal posts and recent interactions are also strong predictors.

1. Introduction

While many different online health communities exist, each having their own purpose and spin on the concept, we focus on CaringBridge, an online health community founded in 1997. CaringBridge is a 501(c)(3) non-profit organization that brings together millions of people annually in an online social network to help overcome the isolation often experienced during a medical journey.

The primary function of CaringBridge is to host personal blogs, known on CaringBridge as *sites*. Each site focuses on the health journey of a single patient, comprised of multiple *journals* posted throughout the health journey. Journals are written by *authors*, who can either be the patient themselves or the caregiver of the patient. Each journal on every site is typically written by the same author, although this does not necessarily need to be true.

A site can define its privacy level- viewable by the public, CaringBridge registered users only, and approved users only. *Visits* are defined as a user simply viewing a journal. CaringBridge also offers site visitors three ways to interact with the site. Each journal has a section for replies, where visitors are able to leave comments regarding the contents of the journal. Additionally, each site has a single Guestbook page, where visitors can leave comments not specifically related to any particular journal post. Finally, visitors can leave "likes" on journals and comments.

The concept of social support motivates parts of our study design. When CaringBridge was created, sites were intended to inform the patient's pre-existing social network about their health journey. However, over time, CaringBridge has become a home for forming entirely new

connections between users. This is important because social support leads to stronger patient outcomes (Holt-Lundstad 2010), and we want to continue designing online health communities to facilitate increased social support among users. For this paper, we limit our user nodes in the interaction network to be authors of sites, not visitors, for two reasons. First, authors typically have more features available for analysis; this can potentially lead to greater and more interesting insights into how connections form. Second, the helper principle states that when an individual provides assistance to another individual, the first individual may benefit as well (Riessman 1965). In addition to measuring health outcomes from the target author as a result of individual connections, we can begin to learn how providing social support on an online health community can increase the initiating author's health outcomes as well.

Using Authors as graph nodes and Interactions as graph edges, we can define CaringBridge itself as a dynamic, directed graph of author interactions. Dynamic graphs refer to networks that vary over time, so the state of the interaction network at time t_0 refers to all of the interactions between users which have taken place before time t_0. Directed graphs are graphs where its edges have defined starting and ending nodes, as opposed to an undirected graph, where edges simply exist between nodes, without a defined direction; this reflects how interactions are initiated by one author and received by the target.

This paper analyzes the growth of the author and interaction network over time. Because each node is a discrete entity, the connections which it initiates can be thought of as individual choices made by the authors. We want to understand the creation of these connections: when they are formed, who they are formed between, and the characteristics of the users involved.

2. Data

Through a research partnership between CaringBridge and the GroupLens Research Lab, this work contains analysis of de-identified data from 588,210 sites and 22,333,379 users created between June 1, 2005 and June 3, 2016, shared in accordance with the CaringBridge Privacy Policy and Terms of Use Agreement. Due to the sensitive information contained in this data, we are not able to publicly release the dataset used for analysis, although any questions may be directed to either the author or CaringBridge directly.

Because this research focuses on interactions between authors, we limit ourselves to analyzing only the 565,857 authors in the dataset. We further limit the dataset to authors with at least 2 posts and an author tenure of at least 24 hours between the first post and the last post in order to remove all non-relevant authors, including spam and sites which are too short-lived to be useful for analysis. In total, this gives us 327,444 authors and 13,081,711 journals. As authors are allowed to author posts on multiple sites, these journals span across 345,463 unique sites.

Looking specifically at interactions, defined previously as either journal replies, guestbook comments, or likes, we find a total of 2,774,494 interactions where both the initiator and recipient were found in the list of non-spam authors.

3. Methodology

In Overgoor et al. (2018), a framework was described for modeling social network formation using discrete choice theory. Discrete choice theory is used to model how individuals make choices between discrete alternatives, where each alternative has different qualities, or

features, which influence the choice made by the individual. In order to model these discrete choices made by individuals in social networks, Overgoor uses *Conditional Multinomial Logistic Regression*, hereafter referred to as a *Conditional Logit Model*, combining multinomial logistic regression models with conditional logistic regression models. Multinomial logistic regression models use input features to predict multi-class output variables. These are necessary for the CaringBridge dataset, and social networks in general, because each individual chooses their next interaction between all other users in the social network (i.e. more than the two options that standard logistic regression would account for). Conditional logistic regression models extend the scope of input features in standard logistic regression from one set of features between all models to features that are independent between the alternative choices. For example, in the CaringBridge network, authors will choose to initiate an interaction not only based on features from the source author, but by features that vary between the targets, such as the target author health condition.

In the actual CaringBridge network, whenever a user initiates an interaction, they are selecting between every single potential target in the network. However, when training our conditional logit models, selecting from the pool of every potential target quickly becomes computationally infeasible because the size of the pool is equal to the number of other authors: 327,444. Instead, we perform a process known as *negative sampling*, to reduce this search space.

Let us demonstrate negative sampling with a hypothetical example. At time t, a base user u_0 initiates a connection with target user t_0. Given this interaction, we want to understand why the target t_0 was chosen instead of a different target from the pool of potential targets $t_1 ... t_k$. With negative sampling, we randomly select n targets from the k-sized target pool who were not

chosen by u_0. In this paper specifically, we are reducing our total pool size of k=327,444, to a small pool with n=24. After including the initial positive sample, the target that the source user actually initiated with (t_0), this gives us 25 potential targets per interaction for the source user to choose between.

Next, we select which features to use to train our model. In Overgoor et al., Flickr connections and a network of research citations were used as sample real-world datasets. For their predictive features, they primarily used features derived directly from the graph network itself: indegree, outdegree, and whether the target is a friend-of-a-friend. Indegree refers to the number of edges directed towards the node under analysis, while outdegree refers to the number of edges originating from the node. Friend-of-a-friend means that the target node can be reached from the source node in two hops.

The first category of features we are implementing are those derived from the interaction graph itself. This interaction graph is a dynamic, temporal network, where nodes represent users and edges represent interactions. Our basic graph features were taken directly from Overgoor et al.: Indegree, Outdegree, Connectedness between base and target user, and Reciprocity between base and target user. Additionally, we create two different graphs - a combined author and visitor interaction graph one containing 25,634,494 and a subset of this graph, an author-only interaction graph, containing 827,053. Note that an author does not need to have authored any posts at the time of the interaction for the interaction to be present in the author-only interaction network. If a user's first post took place after an interaction, the user is still considered an author at the time of the interaction.

For the feature generation of the combined interaction network, we order all interactions by their timestamps. Iterating through these interactions, we sequentially append each interaction to a progressively larger interaction network. Before even computing any features, we first generate approximately 80% of the graph as our starting network, the first 21,578,204 of 25,634,494 interactions. Then, we iteratively add the remaining interactions and randomly sample these iterations as we build out the network, with each interaction having a 10% chance of being added to the list of observations. Due to the massive size of this network, following this iterative process is relatively slow, so the number of interactions in our sample is only 3,740. Then, for each interaction, we perform the negative sampling technique to find 24 other target users that the initiating user could have interacted with instead of the actual target user, which gives us 25 observations per sample and a total of 93,500 total observations.

For the author-only interaction network, we randomly sampled 22,000 authors from the list of 191688 authors with at least one interaction initiated to another user. Using the same negative sampling process as the combined author and visitor interaction network with n=24, we achieved a total sample size of 550,000. Taking advantage of the significantly reduced network size, we precomputed the indegree and outdegree for every user at every possible interaction timestamp. Additionally, calculating interaction reciprocity is very easy, as looking up the existence of the reverse interaction is a quick operation. However, calculating the distance between nodes is inefficient without the actual network itself, so we were forced to forgo this feature for the author-only network.

Beyond the basic graph features, we are more interested in specific features related to our understanding of the social network as an online health community. We selected several features

to implement based on the characteristics of a user. Initial features were selected in a very wide manner, based on nearly all of the datasets provided to our lab by CaringBridge. These features can be broken up into various categories of statistics about the target user: total author activity, recent author activity, self-reported user data such as health condition, and various pre-computed author statistics, such as if the author is a patient or caregiver.

4. Results

4.1. Combined Author and Visitor Interaction Network

In the combined author and visitor interaction network, we have 6 features over 3,740 unique source nodes, and 93,500 total interactions.

	interaction	**output**	**outdeg**	**indeg**	**recip**	**is_connected**	**has_indeg**	**has_outdeg**
interaction	**1.000**	-0.020	-0.007	0.000	-0.005	-0.042	0.016	0.008
output	-0.020	**1.000**	0.128	0.084	0.101	0.491	0.584	-0.603
outdeg	-0.007	0.128	**1.000**	0.254	0.084	-0.014	0.163	0.087
indeg	0.000	0.084	0.254	**1.000**	0.004	0.003	0.117	-0.009
recip	-0.005	0.101	0.084	0.004	**1.000**	0.023	0.071	0.004
is_connected	-0.042	0.491	-0.014	0.003	0.023	**1.000**	0.010	-0.586
has_indeg	0.016	0.584	0.163	0.117	0.071	0.010	**1.000**	-0.444
has_outdeg	0.008	-0.603	0.087	-0.009	0.004	-0.586	-0.444	**1.000**

Table 1: Correlation values between input features across 93,500 observations

In Table 1, we see the correlation between the interaction number, output, and the six input variables. Interaction refers to the enumerated target alternative for a given observation, between 1 and 25. Output is a binary variable, which is 1 if the observation contains a real interaction and 0 if the observation is derived from the negative sampling process. Because of the way our observations are constructed, alternative 1 is always the positive sample, while alternatives 2-25 are derived from the negative sample.

Visualizing this correlation table is important because it helps us determine if any of our input features are overlapping. If the correlation between two input features is too close to 1 or -1, then they are effectively redundant, and one of them can be dropped from our analysis. However, the only relatively high correlations are between *output* and *has_outdegree* (-0.60), between *output* and *has_indeg* (0.58), and between *output* and *is_connected* (0.49) These correlations are not issues because a high correlation between the output and an input feature is desirable, since the goal of our analysis is to predict the output from our input features.

	1. Indegree	**2. Outdegree**	**3. Reciprocity**	**4. Connected**	**5. Combined Degree**	**6. All features**
log(indegree)	-0.364* (0.023)				-0.528* (0.028)	-0.536* (0.028)
has_indegree	6.220* (0.104)				5.994* (0.125)	6.005* (0.127)
log(outdegree)		1.635* (0.031)			1.119* (0.045)	1.101* (0.046)
has_outdegree		-5.413* (0.082)			-1.245* (0.110)	-1.270* (0.112)
is_reciprocal			25.000 (15051.210)			22.632 (2802.773)
is_connected				26.000 (10569.660)		22.506 (3084.050)
Test Accuracy	0.7718	0.6353	0.2765	0.2706	0.8118	0.8265

Table 2: 6 conditional logit model fits for combined author and visitor interaction data. Standard errors of the estimates are given in parentheses. Model was computed over 3,400 training observations and evaluated over 340 test observations. Asterisks indicate that the value is statistically significant with $p<0.01$.

The first two models were trained on indegree features and outdegree features, respectively. Both models consist of a simple binary variable describing if the degree exists (degree > 0), as well as log(degree). Taking the log of the degree is required because of the huge disparity in degrees from author to author; the log function reduces these order-of-magnitude differences into more manageable output for the conditional logit model. The indegree model has an accuracy of 0.76, and the outdegree model has an accuracy of 0.64. Compared to the dummy classifier accuracy of 1/25, or 0.04, these accuracies are very high, indicating that the indegree and outdegree by themselves are strong predictors of interaction alternative choice.

When interpreting the model coefficient values, the most important element to note is the sign of the coefficients. A positive sign indicates that if the value of the observation's feature is increased, the observation has a higher likelihood of being chosen, while a negative sign indicates a lower likelihood of being chosen. Thus, we see that having an observation with indegree > 0 causes a feature to have a higher likelihood of being chosen than an observation with indegree=0, but each additional incoming edge decreases the likelihood of an observation being chosen.The reverse is true for outdegree, with has_outdegree having a negative correlation and log(outdegree) having a positive correlation.

Our next two models are the reciprocity of a connection, and if the source node and observation are distantly connected in the graph. We would expect these features to be strong predictors of interaction choice, since intuitively, people connect with others closer in their social network than others farther away. In fact, we do see this with overall accuracies of 0.28 for reciprocity and 0.27 for is_connected, as well as positive coefficient values for both features.

However, we cannot read into these values too deeply, because the coefficient values are not statistically significant.

The final two models are simply two possible combinations of the previous models - combined indegree and outdegree, and the combination of all features. Combining indegree and outdegree performs better than either indegree or outdegree alone, with an accuracy of 0.81. Combining all features together performs even better, with an accuracy of 0.83.

4.2. Author-only Interaction Network

In the author-only interaction network, we have 32 features over 22,000 unique source nodes, and 550,000 total observations.

The full correlation matrix is too large to be shown below. However, the only significant correlation values (either greater than 0.5 or less than -0.5) were between num_total_interactions and indegree (0.54), time_since_first_post and time_since_recent_post (0.95), proportion_finished_journals and has_finished (0.72). Descriptions of these variables can be found further into the results sections. Of these correlations, only time_since_first_post and time_since_recent_post is excessively large, so in the future, we may not need to compute both these values.

	1. Degree	2. Condition	3. Authorship	4. Author Interactions	5. Author Journals	6. Recent Interactions	7. Reciprocity	8. All Features
log(indegree)	0.177* (0.007)							0.293* (0.011)
has_indegree	-0.388* (0.031)							-0.669* (0.047)
log(outdegree)	0.076* (0.011)							0.182* (0.018)
has_outdegree	-0.069* (0.019)							-0.340* (0.028)
is_shared_condition		0.0669* (0.043)						0.165* (0.062)
cond_custom		-0.841* (0.032)						-0.034 (0.045)
cond_Neurological		-0.734* (0.066)						-0.076 (0.092)
cond_Injury		-0.608* (0.054)						0.061 (0.074)
cond_Cancer		-0.767* (0.054)						-0.067 (0.034)
cond_Surgery		-0.658* (0.050)						-0.010 (0.069)
cond_Unknown		-0.738* (0.107)						-0.091 (0.146)
cond_Childbirth		-0.574* (0.064)						0.004 (0.088)
cond_Other		-0.617* (0.092)						0.025 (0.127)
cond_Congenital		-0.584* (0.141)						0.166 (0.187)
cond_Cardiovascular		-0.554* (0.056)						0.046 (0.078)
is_patient_authored			0.000 (1.10e14)					
is_mixed_authored			0.000 (6.83e13)					
is_shared_authored			0.000 (4.34e13					
visit_count				-0.00001* (0.000)				0.00000 (0.00000)
num_total_interactions				0.0002* (0.000)				0.00004* (0.00001)
num_previous_replies				0.035* (0.002)				0.008* (0.002)
time_since_first_post					-0.002* (0.000)			-0.001* (0.00004)

num_prev_journals					0.003* (0.000)			-0.001* (0.0002)
proportion_finished_journals					-3.327* (0.036)			-2.980* (0.047)
has_finished					-2.198 (0.044)	0.00001* (0.000)		-2.094* (0.053)
time_since_recent_post						0.524* (0.004)		0.003* (0.00001)
num_interactions_prev_week						0.035* (0.001)		0.246 (0.004)
num_posts_prev_week							9.855* (0.500)	0.008 (0.001)
is_reciprocal								10.744* (0.610)
Test Accuracy	0.0865	0.054	0.039	0.1185	0.5565	0.4955	0.1999	0.6815

Table 3: 8 conditional logit model fits for author-only interaction data. Standard errors of the estimates are given in parentheses. Model was computed over 20,000 training observations and evaluated over 2,000 test observations. Asterisks indicate that the value is statistically significant with $p<0.01$.

Our first model is the combination of indegree and outdegree. However, unlike in the combined author and visitor interaction graph, these indegree and outdegree features refer only to prior interactions made with other authors. Somewhat unexpectedly, the results of this model heavily contrast with the first indegree model, with an accuracy of only 0.09, compared to 0.81 in the combined author and visitor degree model. Additionally, the coefficient signs for indegree are flipped, with a positive coefficient for log(indegree), but a negative coefficient for has_indegree. Further analysis is needed to determine why the indegree and outdegree signs are not consistent with each other between the author-only graph and the combined author visitor graph.

The second model is based on a categorical feature, the self-disclosed condition of the author. On CaringBridge, authors are able to display one of seven different conditions on their site, such as Cancer or Childbirth. Additionally, options for “Other condition”, “Condition

Unknown", or a custom condition are offered. In order to avoid the dummy variable trap for categorical variables, we exclude the feature where the user has not disclosed their condition. We also include a computed feature, is_shared_condition, which is 1 if both the source author and target author share the same condition, and 0 if either of the authors have not disclosed their condition, or if the conditions are not the same. We would especially expect is_shared_condition to be a good predictor of interaction choice. However, even though the coefficient sign for is_shared_condition is positive, the accuracy of this model is low, only 0.05, barely better than the dummy classifier.

The third model is based on another categorical feature - authorship type. On CaringBridge, journal posts on a site may be written by either the patient themselves or a caregiver. In previous work with the CaringBridge dataset, our team built a machine learning classifier to determine if the author of any given journal post was a patient or caregiver. We then define an author to be a patient if at least 90% of all posts written by an author were classified as patient, and a caregiver if less than 10% of posts written by the author were classified as a patient. Otherwise, we classify the author as a mixed-author. Like with the condition variable, we exclude one of the variable options, is_caregiver_authored, to avoid the dummy variable trap. Additionally, we include another computed feature, is_shared_authored, which is 1 both the source author and target author share the same authorship type, and 0 otherwise. However, the results of this model are low, only 0.04, the same as the dummy classifier, and none of the feature coefficient signs are statistically significant.

The fourth model is based on previous target author interactions - total number of visits, number of total interactions (both incoming and outgoing), and number of prior replies on the

author's journals. While the sign of the feature coefficient of visit_count is negative, num_total_interactions, and num_prev_replies are positive, and the model's overall accuracy is 0.11, somewhat better than the performance of the dummy classifier.

The fifth model is based on the target author's prior journal posts - the amount of time since the author's first post, number of previous posts, the proportion of posts authored at the time of interaction relative to the authors total number of posts, and if the author has finished writing journals (i.e. the author will not author another post after the time of interaction). Time_since_first_post, proportion_finished_journals, and num_previous_journals have negative coefficient signs, while num_prev_journals have positive coefficient signs. Overall, the accuracy of this model is 0.56, significantly better than the dummy classifier's accuracy of 0.04.

The sixth model is based on recent journals and interactions of the target user, where we define "recent" as an event occuring in the week preceding the interaction. These features include the time since the most recent post, the number of total interactions in the previous week, and the number of total posts authored in the previous week. All three of these features have positive coefficient signs, and the total accuracy of this model is 0.50, also much better than the dummy classifier.

The seventh model is based on reciprocity, similar to the reciprocity in the combined author and visitor graph. As in the earlier model, reciprocity here has a positive coefficient sign, but the sign's positivity is statistically significant here. However, the overall accuracy is 0.20, higher than the dummy classifier, but lower than the reciprocity accuracy of 0.28 in the combined author and visitor graph.

Our final model is the combination of all the prior seven models, and the accuracy of this model is 0.68. As in the combined author and visitor graph final model, the accuracy is higher than any of the individual models.

5. Conclusion

During this study, we created 14 conditional logit models to better understand the formation and growth of interactions in an online health community. While the specific results from our models are preliminary, we can make two broad types of comparisons: between the combined author and visitor graph and author-only graph, and between the purely-graph features and features specific to online health communities.

First, we see that adding interactions with non-author visitors greatly improves the accuracy of the overall results, even when only modeling author-author interactions. Specifically, adding visitors to our degree models improves accuracy from 0.09 to 0.81 and increases the total combined model's accuracy from 0.68 to 0.83. Second, we see that several of the health community-specific features, such as recent interactions and prior journal posts, do improve the accuracy of models. In fact, they account for the majority of accuracy for the author-only combined model, as the recent interaction model has an accuracy of 0.50 and the prior journal post model has an accuracy of 0.56, while the combined model has an accuracy of 0.68.

Although computational infeasible for this study, in future work, we would like to combine the interaction features with the online health community features. Additionally, we

would like to create new features derived from the online health community journal posts, such as sentiment analysis of journal posts and mortality outcomes for patients.

Overall, we have found that the framework of conditional logit models for modeling network growth described in Overgoor et al. can be successfully extended to the domain of online health communities. Our major next steps are expanding and refining our analysis of these conditional logit models to broaden our understanding of online health communities as a whole.

References

[1] Holt-Lunstad J, Smith TB, Layton JB (2010) Social Relationships and Mortality Risk: A Meta-analytic Review. PLoS Med 7(7): e1000316. https://doi.org/10.1371/journal.pmed.1000316

[2] Neal, Lisa; et al. (2006). "Online Health Communities". CHI'06 Conference on Human Factors in Computing Systems. Montréal, Québec, Canada: ACM.

[3] Overgoor, Jan & R. Benson, Austin & Ugander, Johan. (2018). Choosing to grow a graph: Modeling network formation as discrete choice.

[4] Riessman, F. (1965). The "helper" therapy principle. Social Work, 10(2), 27-32.