# Predicting User Compatibility Online with Advanced Natural Language and Network Features

*By Jay Hack, Sam Beder, and Alex Zamoshchin*

## 1. PROBLEM STATEMENT

*Chatous* is a recently-developed social network that pairs semi-anonymous users for random, IM-style chats. It has become relatively popular since its inception a year ago, with typical traffic of several thousand users at any given time. One of the foremost challenges facing the site is how to match users in order to maximize their individual enjoyment. While the full user-matching algorithm can be stated (and is currently stated) as a search problem, where states are pairings of users, an integral component of this algorithm is the set of heuristics used to predict the compatibility of two users given limited information on them. In this paper, we address the following question: how might one accurately predict the compatibility of two *Chatous* users? That is, concretely, given two users, what is the best method for predicting the expected chat length between them.

## 2. PRIOR WORK

Our primary inspiration is the initial analysis of the networks and interactions on *Chatous* performed by Guo et al. In their paper, they introduce *Chatous* and provide high-level analysis on its properties. Furthermore, they provide the reader with a description and evaluation of an algorithm they developed to predict unknown edge weights. Guo et al. define user compatibility between two users as the geometric mean between their respective ratings of a shared online interaction. They employ a linear classifier trained on a portion of their data set; then, in order to evaluate their classifier's performance, they measure the percentage of the time that it correctly predicts which of two unobserved edges is of higher weight. In their paper, they claim a 93% success rate.

As we will discuss in our paper, it will be difficult to directly compare our results to that of the initial *Chatous* paper. Through our own finding and discussions with Guo et al., many of their initial results have been invalidated as improper measurements of success. Their triad metric in particular, which is their strongest measure of success in their published work, is argued to not be significantly predictive on real data. As a result, our measures of success will rely on both our own metrics and those used in the original *Chatous* model, but we will not implement some of the specific features of this prior work, as they have shown to be weak.

While Guo et al.'s work applies directly to our chosen task, we also plan to use the findings of two other papers describing similar approaches to signed edge classification in different domains. Leskovec et al. examine various predictive techniques in social networks composed of 'positive' and 'negative' edges. In particular, they describe an edge-weight classifier trained on features of neighboring relationships and neighboring-neighboring relationships, a model based upon theories of balance and status, and an approach to edge-weight prediction using heuristics. Chiang et al. also explore sign prediction, though they focus on the insights offered by of longer 'walks' about the network. Their treatment of longer walks can be considered an expansion upon or generalization of the idea of triads.

## 3. METHODS

In this project, we developed a set of feature-extractors and classifiers that demonstrated high accuracy in predicting the length of chats between unknown users. As a baseline, we made use of a set of features that described objective aspects of a pair of users (age difference, etc.). Our principal contributions were (1) the development of a set of network-related features and (2) the application of unsupervised feature learning, such as Latent Dirichlet Allocation on chat content and Random Forests on general features, in order to better characterize users. We describe all of these feature sets below.

## 4. DATASET

Our dataset consists of metadata describing approximately 9 million conversations occurring between 80,000 users on *Chatous*, as well as the

following information on the individual users: Geographic location, age, gender and unigrams from a short personal statement. (The unigrams were hashed in the interest of user privacy.) In addition, for each conversation, we have access to a bag of words representation of the users' dialogue. We modelled this dataset as a graph, with users represented as nodes and chats represented as weighted edges in which the weight corresponded to the number of lines of the chat.

Due to the relatively recent founding of *Chatous*, this graph is extremely sparse (See *Figure 1*). Furthermore, the set of chats is heavily skewed towards very short interactions, with zero-length chats making up the majority of all of them (See *Figure 2*).

## 5. BASELINE FEATURES

As a baseline we introduce a number of features directly relevant to a specific chat. Such features were introduced in order to identify any correlations between simple chat metadata and chat success. *Boolean Gender Equivalence* was a binary variable indicating whether the two participants were of the same or different gender. *Age Difference* was the absolute value of the difference in age between the two users. *Fraction of Disconnects* was the absolute value of the difference in the proportions of conversations disconnected by each user. Finally, *Fraction of Lines in Conversation* was the absolute value of the difference in the proportions of lines sent to total lines for each user. Together, these four variables were used as baseline features in the classification of user compatibility.

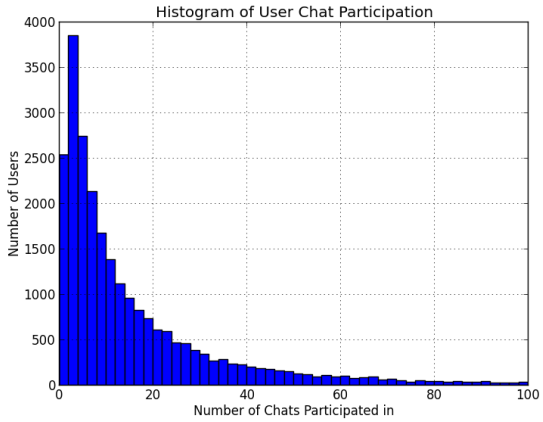


**Figure 1: Degree distribution of the *Chatous* network**

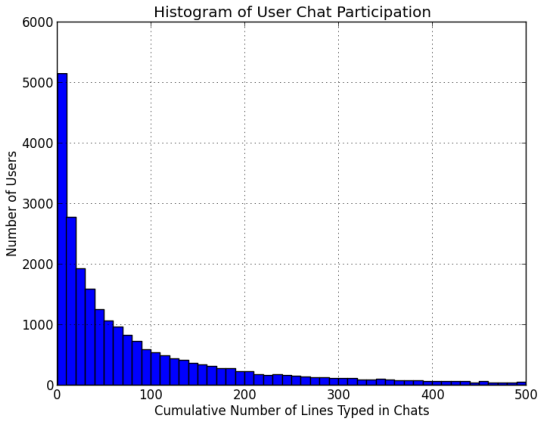


**Figure 2: Histogram of user chat participation**

## 6. TOPIC FEATURES

Perhaps our most effective and original contribution was our application of topic modeling, namely Latent Dirichlet Allocation (LDA), in order to discern what 'topics' a given user was most interested in or likely to react well to. At a high level, we first found the topics that each chat was composed of, then used these results to determine what topics any given user was most likely to engage in or react favorably to. Having associated each user with their set of topics, we computed feature vectors for pairs of users based on the difference in their preferences in topics. Below is a description of LDA and a more precise specification of how we computed the topic features for a pair of users.

### *LDA Description*

LDA is a generative probabilistic model that models each document in corpus as a being generated from a finite mixture of underlying topics. Given a corpus of documents and a number of topics to model, LDA utilizes the EM algorithm to infer distributions over possible values for hidden (latent) variables governing the documents' generation. In particular, for each document, it produces a probability distribution over the set of possible topics from which the document has been generated; in addition, it finds prior probabilities for each topic appearing in any document and, for each topic $z_i$, a distribution over words $p(w_i|z_j)$ representing the probability of word $w_i$ being generated in a discourse on topic $z_j$. See the

Bayesian template model representation below.

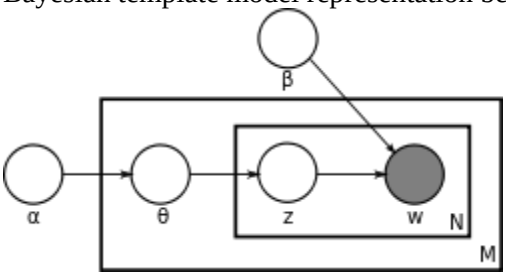


**Figure 3: Template model representation for LDA**

Here, the outer plate represents individual documents (of which there are $M$) and the inner plate represents individual words occurring within a given document (of which there are $N$). $w$, the words themselves, are the only observed variables, while $z$, loosely interpreted as the topic that the creator of the document had in mind when he wrote word $w$, is latent, as well as $\theta$, a distribution over the topics $z$ contained in the current document. The other parameters, also latent, are hyperparameters governing the prior probability of seeing a given theta and the prior of seeing a given word.

The following is a full generative story of how a document is created, according to LDA:

- Choose M ~ Poisson($\xi$) for some $\xi$ (the number of words in the document)
- Choose $\theta \sim \text{Dir}(\alpha)$
- For $i = 1 \ldots N$
    - Choose a topic $z_i$ ~ Multinomial ($\theta$)
    - Choose a word $w_i$ from $p(w|z_n; \beta)$, a multinomial probability dist. conditioned on the topic $z_i$.

Thus, this generative Bayesian model is capable of characterizing individual documents' distribution over possible topics $z_i$ via the latent variable $\theta$. Our application used the open-source python package Gensim to find $\theta$, henceforth referred to as 'topic vector,' for each document.

*Features from LDA*

Having found the topic vector $(i)$ for each chat $d(i)$, we then computed for each user a 'weighted topic vector,' or a weighted average of the topic vectors over all of their chats, in order to characterize that user's preference in topics. Then, for any combination of two users which we tried to predict compatibility, we would compute a set of features consisting of the element-wise difference between their respective feature vectors. These will henceforth be known as 'topic features.'

## 6. NETWORK FEATURES

Several network features contributed greatly to the success of our overall algorithm. Further, these features gave us the greatest insight into the structure of the *Chatous* network. From a network perspective, the *Chatous* conversation graph is particularly interesting compared with other well-studied networks, and presents some unique challenges. Since the edges are constructed by the *Chatous* algorithm and not by users, we cannot infer results simply based on the existence of an edge or path between nodes. Furthermore, as we see in *Figure 2* the large majority of edges have a chat length of zero. Therefore trying to calculate predictions based only on these edge values quickly loses its predictive power as the vast majority of these chats are failures and provide no insight into user preference. As a result, we compiles an assortment of features that balance restricting our edges to those that give strong indication of preference and the use of a large set of paths to combat the graph sparsity.

Our first feature finds the optimal path score between two given nodes based on a metric using the chat length of edges. We first restrict our network to only contain edges of at least a certain chat length, and then use these more sparse graphs to find the largest edge restriction that still allows for a path between the two edges. Formally, we calculate this metric by taking

$$\max_{p \in P} \min_{e \in p} length(e)$$

where $P$ is the set of paths connecting $u$ to $v$ for each pair of nodes $u$ and $v$. This feature attempts to combat the large majority of chats that are of length zero. Since these chats are such a large part of the dataset, they make it difficult to predict when two given nodes will have a long, positive conversation. Therefore, the existence of a path between these nodes that consists of only edges with long conversations should be a positive indication of a success between the

endpoints. This metric finds the path that that the largest conversation length such that all edges along that path will be at least that given length, resulting in a larger score for nodes that have strongly positive paths between them.

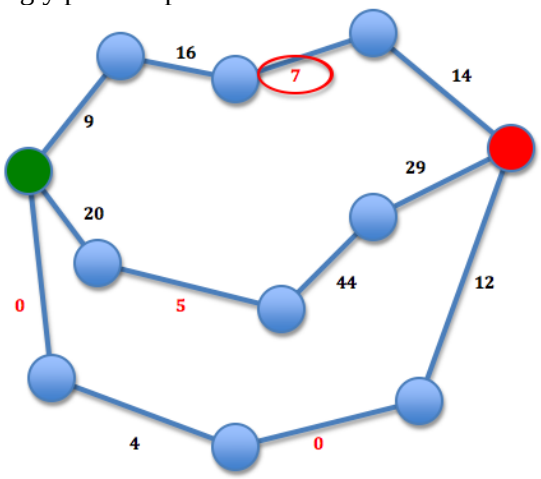


**Figure 4: Our first metric chooses a score based on the maximum chat length over the minimum length of a chat along all paths between our two nodes. Here, the top path has the best path, returning a score of 7 for these two nodes.**

Our next network feature used a combination of long walks between two given nodes. It first calculates all path between the two nodes under a certain path length threshold. It next calculates score by taking the average over all paths, where the score of a path is calculated by:

$$\Sigma_{p\in P}\,\Pi_{e\in P}\left(1-\left(\frac{1}{length(e)+1}\right)\right)$$

where $P$ is the set of paths connecting $u$ to $v$ for each pair of nodes $u$ and $v$. This formula gives us the desired results for several reasons. In graphs that are sparse, single paths may not provide much insight, but many long paths can allow for predictive power when used together (Chiang et al.). Our formulation uses the set of long path and produces a score that both rewards shorter paths and paths with mostly chats of high length. This first point is desirable because in longer paths we lose insight between the two endpoints as the length of the path grows. The second point is equally desirable because as we have seen above that there are a large number of chats of zero lengths, which provide little information. Therefore we can tell much more about the endpoints of a path when the edges along that path have larger chat lengths. Finally by taking the average of these path scores we come to a cumulative score that provides deeper understanding of the relationship between the two desired nodes.

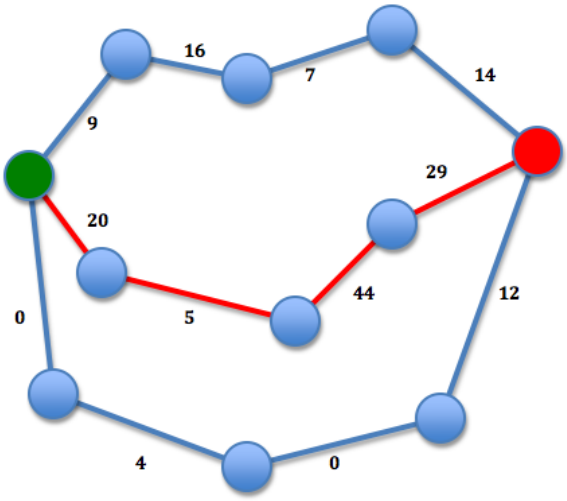


**Figure 5: Our second metric calculates a score based on the product of a function of the edge lengths along long walks in the graph. Here, the middle path maximizes this product, and will make the largest contribution to the average.**

While our first network feature finds the best single path in a restricted graph, this second feature takes an average over all long paths. Therefore we are utilizing disparate and uniquely insightful aspects of our *Chatous* network. These features when taken in combination provide a measure the closeness of the two nodes with the best given path, and the overall relationship between the nodes with our averaging of long paths.

## 7. INITIAL FINDINGS

### *Insights on Weighted Topic Vectors: Personality Clusters*

In order to demonstrate to ourselves that the weighted topic vectors actually provided insight into a user's character and/or their tendency to form positive relationships with other users of certain types, we produced a set of visualizations describing these vectors. This consisted of clustering users based on their weighted topic vectors and then modelling the compatibility between said clusters.

*Personality Clusters*

Intuitively, we reasoned that users' weighted topic vectors would not be generated completely independent of one another; that is, it seems as though there are a set number of personality types from which each user is generated, and a user's weighted topic vectors is largely dependent on their personality type. In order to gather support for this idea, we produced a two-dimensional visualization of 5000 of the users' weighted topic vectors using t-SNE, shown below in *Figure 7*. One can see that while the center forms one large, convoluted cluster, there is a multitude of small, dense clusters containing between 5 and 50 users surrounding the massive center cluster. We interpreted this as evidence that such 'personalities' really do exist.

*Compatibility between Clusters*

Having determined that such personalities probably do exist, we performed k-means clustering on all users' weighted topic vectors in order to assign them to a personality cluster, then found the average chat length between users of each cluster. (See a heat map of our results in *Figure 6*) As one can see, these personality clusters are clearly meaningful, as the attraction between different personality clusters varies widely. Another insight that we gained from this is that *users are not particularly attracted to other users of precisely the same interests*. If this were the case, then the diagonal of *Figure 6* would be much 'hotter,' whereas currently the 'hottest' elements in the graph seem to be somewhat arbitrarily chosen between the clusters.

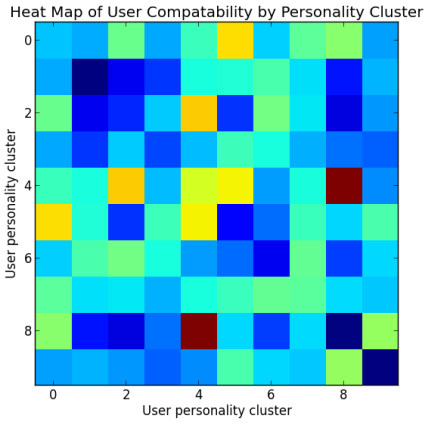


**Figure 6: Heat map of user compatibility by personality cluster. The $(i, j)^{th}$ square in the matrix represents the average compatible between the $i$ and $j^{th}$ personalities.**

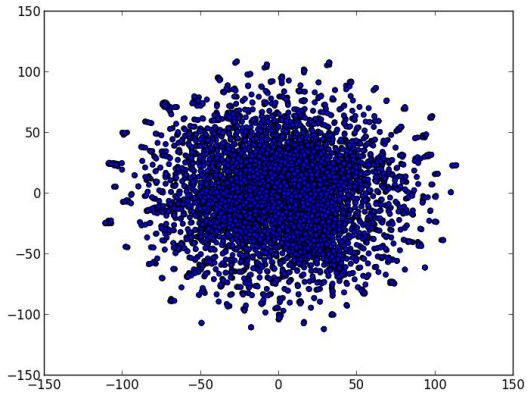


**Figure 7: 2-dimensional visualization of the first 5000 users' weighted topic vectors, produced using t-SNE**

*Network Insights*

In addition to providing valuation metric towards predicting node compatibility, our network features also provided large insight into the structure of the *Chatous* network. We provide two main metrics, a notion of the best path between nodes and an average of long walks between paths. While we found that this second feature gave us the best predictive power, the first metric shows some interesting results about the overall network. We see in *Figure 8*. a histogram of our first feature between all pairs of matched nodes. We see that in the full dataset that there almost always exists a path between any pair of nodes. Even though a huge portion of our conversations are length zero, we note that most edges are still connected when we restrict their connections to paths containing only conversations of at least moderate length (length greater than four lines). The drop-off thereafter is rather steep, and we observe that there are very few pairs of nodes connected by a path of conversations that are longer than ten lines.

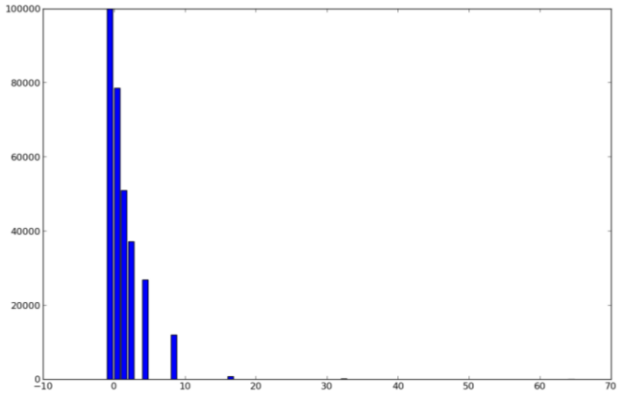

**Figure 8: A histogram showing our best path network feature between all matched nodes, placed into bins of powers of two for computational efficiency. The first bar requires no path between the pair (showing that there is a total of 100,000 pairs in question), and each subsequent bar shows paths of conversations of higher chat length.**

## 8. EVALUATION METRIC

For evaluation, we use a method similar to that of Guo et al.: our success is precisely the percentage of the time that we correctly classify which of two unknown edges adjacent to a given node is larger. (This can be interpreted as a heuristic for which of two users a given third user should be matched with.) More specifically, we take a random sample of 1000 users who have had a chat with at least two users. For each user, we randomly select two of their chats, and use our model with the two chats held-out to predict a total number of lines for each chat. Using these two predictions, we select the chat with the highest predicted length as the best match for the given user. We then use the actual conversation length to determine the actual best match and classify our prediction as correct or incorrect. If there is no best match (when both chats have the same length), we skip this user since any model is guaranteed to classify correctly, thus limiting our results to only those choices with a clear winner. Finally, we calculate the percentage of matches performed correctly.

We compare to prior work by Guo et al. and see how our methods compare to that of the founders of *Chatous*. However, as mentioned above, flaws in their conclusions make any direct comparison difficult. We rely on a comparison to a random classifier, and attribute any improvement over 0.5 to predictive power in our model.

## 9. RESULTS

We use baseline, topic, and network features and attempt to optimize for the evaluation metric discussed above. Finally, we use a set of all features in an effort to create an ensemble with strong predictive power. We analyze a variety of methods, including different types of regressors and label functions, in an effort to perform the best matches as determined by our evaluation function.

### *Linear Regression, K-Nearest Neighbors, and Random Forest*

We began our analysis through use of Linear Regression as a simple and effective model. As seen in *Table 1*, Linear Regression outperforms a random classifier for all three sets of features. More importantly, Linear Regression outputs coefficients, allowing insight into the strength of features we selected. Though t-tests performed on the coefficients of the baseline features set do not indicate a strong correlation, we attribute this to a likely non-linear relationship between the baseline features and the response. More importantly, coefficients both for topic features and network features provide interesting results. For topic features the coefficients represent a variety of values from $-1$ to 1, indicating the relative indicative power of some of the variables, and for network features we see coefficients of $-0.0446$ for the first feature and $0.4505$ for the second. Although this may indicate that no correlation exists for our first network feature, our second network feature was successful.

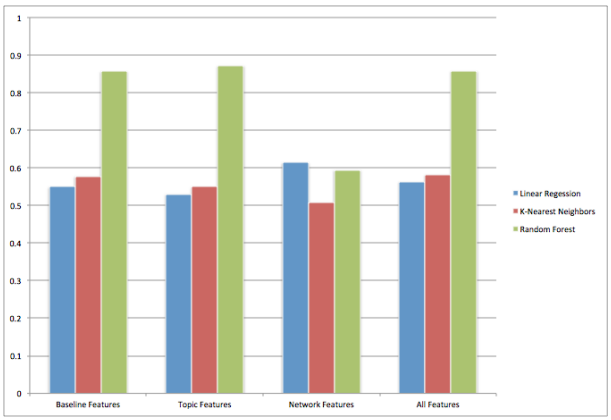


**Figure 9: Prediction Accuracies Using Three Different Algorithms**

| | Baseline Features | Topic Features | Network Features | All Features |
|---|---|---|---|---|
| Linear Regression | 0.549634 | 0.530031 | 0.614155 | 0.562955 |
| K-Nearest Neighbors | 0.577613 | 0.549788 | 0.507462 | 0.58137 |
| Random Forest | 0.856844 | 0.871607 | 0.594339 | 0.857142 |

**Table 1: Prediction Accuracies Using Three Different Algorithms**

Exploration of non-parametric methods, including K-Nearest Neighbors and a Random Forest Classifier, show further improvement. As seen in *Table 1*, through K-Nearest Neighbors produces results similar to that of linear regression, Random Forest significantly outperforms the rest. We attribute the random forest's success to the fact that it learns higher-level representations of its input, and thus has the potential to capture and exploit more of the latent structure present in our data. Although the Random Forest for network features lags behind the other feature sets in its performance, this can be attributed to the fact that it is a model based upon only two features, while all other feature sets contain more. In summary, the Random Forest using all available features achieves a prediction accuracy of 85.7%, a strong improvement both over the random classifier, and all valid results presented by Guo et al.

*Chat length, Log of Chat Length, and Log of Chat Length Buckets*

We next attempt use of a variety of label functions in an effort to help improve selection of the best chat between two. The intuition behind this may be that the problem of prediction of exact chat length is too difficult for simple regression. We consider three methods: the total number of lines in a chat, the logarithm of the number of lines in a chat, and the base 2 logarithm bucket corresponding to the number of lines in a chat.

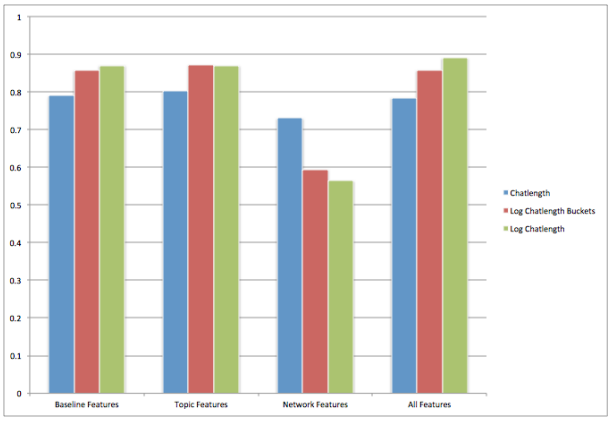


**Figure 10: Prediction Accuracies Using Three Different Label Functions**

| | Baseline Features | Topic Features | Network Features | All Features |
|---|---|---|---|---|
| Chat Length | 0.789802 | 0.801435 | 0.73015 | 0.78435 |
| Log Chat Length Buckets | 0.856844 | 0.871607 | 0.594339 | 0.857142 |
| Log Chat Length | 0.870389 | 0.86970 | 0.56440 | 0.890063 |

**Table 2: Prediction Accuracies Using Three Different Label Functions**

As seen in *Table 2*, direct prediction from chat length proves the most difficult problem, as expected. In fact, this uncertainty in the prediction of chat length directly leads to a higher proportion of incorrectly chosen chats in the best match evaluation metric. Although logarithmic chat length and bucketed logarithmic chat length both prove relatively successful, surprisingly the best results are achieved by a simple logarithmic transformation of the chat lengths. We explain superiority over the bucketed version in that even though the bucketed version provides a more reasonable regression problem for the Random Forest algorithm selected above, a bucketed version is much more likely to produce ties in the selection of the best chat between two. These ties represent a loss of information, since in such a situation the model must randomly select between the two, and a non-bucketed version could have provided information, no matter how insignificant, in order to make the decision non-random. The Random Forest run on the logarithmic transformation of the response variable using all discussed features achieved a prediction accuracy of 89.0%, which represent our best result. This result shows that our combination of features gives us an extremely successful metric in predicting the success of any given chat on *Chatous*.

## 10. CONCLUSION

From the analysis above we see that we were able to achieve a model with prediction accuracy much greater than that of a random classifier and, furthermore, all valid results provided by Guo et al.. Although predicting conversation success in the *Chatous* network is a difficult problem, our best model was able to achieve an approximately 89% success rate in picking the best conversation between two. Such a result is strongly indicative of

the vast potential of more sophisticated natural language and network-related features for the *Chatous* network.

In particular, we see that the most interesting features proposed by our model were the topic clustering created by LDA and the more complicated network-related features than previously considered. In fact, we see that both sets of features are individually powerful and combine to create an even stronger predictor of best match. This indicates that the two sets of features provide different yet complimentary insights on the data.

Suggestions for future work include exploration of other algorithms and more interesting feature bootstrapping and selection. In particular, although removal of the first network-related feature, mentioned above to be likely un-indicative, may result in an increase in predictive power (from decreased noise), further exploration of this feature may result in a valuable modification which aligns the feature more closely to the expected intuition. Moreover, optimization of parameters such as the number of topics constructed by LDA, or the number of weak learners used by the Random Forest, may results in large gains of predictive power, since such optimization has not yet been considered. Finally, exploration of other methods similar to Random Forest such as bagging and boosting, may provide even greater results. Although we have achieved large success, the inherent complexity present in the *Chatous* network means that deeper insights into the network structure could provide even stronger results. We hope that our success with advanced natural language and network-related features will contribute to the role *Chatous* plays in addressing the problem of human compatibility.

## 11. REFERENCES


Antal, T., Krapivsky, P. L., & Redner, S. (2005). Dynamics of social balance on networks. *Physical Review E*, *72*(3), 036121.

Guo, Kevin et al., Predicting Human Compatibility in Online Chat Networks, 2012.

K. Chiang, I. S. Dhillon, N. Natarajan, and A. Tewari. Exploiting Longer Walks for Link Prediction in Signed Network. In Proc. CIKM, 2011.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010, April). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web* (pp. 641-650). ACM.

Mossel, E., & Roch, S. (2007, June). On the submodularity of influence in social networks. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing* (pp. 128-134). ACM.