Link Prediction Analysis in the Wikipedia Collaboration Graph

Ferenc Molnár
Department of Physics, Applied Physics, and Astronomy
Rensselaer Polytechnic Institute
Troy, New York, 12180
Email: molnaf@rpi.edu

***Abstract*—Using the page editing records going back to the very beginning of Wikipedia, we define a dynamic collaboration graph of editors and social links between them. We focus on prediction of social link formation among Wikipedia editors. We show the statistical analysis of five link prediction models, using well-defined statistical measures, such as precision, accuracy, sensitivity and specificity. Results show that the best predictor for screening purposes (identifying most link formations correctly) is given by a model considering the strength of links already existing between the common neighbors of two editors, but the highest probability of correct predictions is achieved by the Adamic/Adar predictor.**

I. Introduction

In the past decade we have witnessed the rapid expanse of the Internet, in both user numbers and volume of contents. It also involved the formation of web-based social networks, which grew to larger sizes than any other social networks before the era of the Internet. One of these social networks was formed by Wikipedia, due to its open editing policy. Editors of certain pages share a common interest in the field that the page belongs to, which is a basis for social links to form between them, and by helping each other, or even competing in perfecting a certain page, social links between them become even stronger.

One of the many exciting questions about these networks is their evolution. It is unquestionably a complex process, and apart from a few general principles (e.g. triadic closure, [2]) the exact dynamics may be different for each social network on the web. Can we find a model that can correctly describes the growth of a social network? Can we predict future links, based on the present state of the network? This is the essence of the link prediction problem [9].

The aim of this paper is to present a thorough statistical analysis of multiple link prediction methods, and find the one that describes the evolution of social links between the editors of Wikipedia most correctly. First, we define the *collaboration graph* as a dynamic graph, constructed from the changelog of Wikipedia over the first ten years of its existence. Then, we give an overview of statistical analysis of binary predictors that we applied in this study. In section III, the results of prediction analysis are presented. In the last section, the results are summarized and compared to each other.

A. Related work

Link prediction belongs to the field of network evolution models, which involves the study of many different social networks, such as citation networks, communication networks, acquaintance networks, and of course, collaboration networks. All of these are strongly linked to the Internet, whose growth and its scale-free degree distribution is well described by the preferential attachment model [1]. It has been shown, however, that the evolution of social networks is driven by a different process [3]. Clustering, also known as the principle of triadic closure [2], plays a very important role.

Collaboration networks have been modelled by their general properties (e.g. [8]), but the problem of precise link prediction was introduced by Nowell and Kleiberg [9], who provide a baseline of link prediction methods, and their analysis. They show that numerous link prediction methods can be significantly more precise, than a random guess. Their work motivated the study presented in this paper. Recently, many models and methods of link prediction were formulated and analyzed (e.g. [6], [7]). The incorporation of time-dependent information to enhance predictions also have gained considerable attention ([5], [4]).

Here, we conduct a thorough statistical analysis of a number of link prediction models, showing the relation between sensitivity, specificity, and accuracy. In addition, the collaboration graph is defined as a dynamic graph, in recognition of the inherent dynamic nature of social networks. Wikipedia is an ideal subject of study, because of the large number of its editors, and its long editing history provides a sufficiently large dataset for sampling and evaluating predictions.

II. Preliminaries

A. Dynamic Collaboration Graph

The collaboration graph, also known as the coeditors graph, a specific kind of social network. Generally, it is defined as the graph composed of editors as nodes; the strength of a link between two editors indicate how many publications (Wikipedia pages, in our study) they edited together in a given timespan. This graph is usually defined as a static graph at a given time, accumulating editing records with some timespan, and social links are inferred from it by setting a threshold on the link strength (i.e. at least how many pages the editors had to edit together).

In case of Wikipedia, we have data for over ten years, with snapshots accumulating the page editing records over weeks. However, instead of using these snapsots as individual static graphs, we join them into a single, large *dynamic collaboration graph*. The nodes are the editors who ever edited a page during the timespan of the entire dataset. The link strengths between nodes change over time, with the same time resolution as the snapshots of the input data, i.e. weeks. The dynamics are defined by the following update rules, evaluated at every time step:

- if two editors edit a page together, strength between them increases by 8.
- every link strength is reduced by 1, until they reach zero.

Using this definition, we can maintain a fine-detailed description of social links between editors. If two editors work together randomly, the link between them is weak, and drops to zero in eight weeks. However, if they are working together repeatedly, the link strength between them keeps increasing, indicating a strong social interaction. In short, we maintain a time-dependent information of past history between editors, not only weekly snapshots, which enables a more precise link prediction for the future.

The dynamics are chosen mainly on the basis of computation efficienty, because the run time of algorithms that generate link predictions strongly depend on the graph being sparse or not, and because our understanding of how the human brain stores (and forgets) long-time memories, including the ones related to social links, is very limited. Exponential decay was also considered, but the problem is that it would excessively prolong the existence of weak links. The graph could become dense over some time, slowing down the analysis so much that it would become unfeasible. However, a linear decay keeps the graph size in check, because links can decay to zero in finite time, at which point they are actually removed from the graph. The strength increment of 8 per editing together is somewhat arbitrary. Based on preliminary computations, random links (no actual social interactions) tend to decay to zero, and true social links also form regardless of strength increment value. Higher increment would only give longer decay times for random links, and higher strength values for social links, but that would only shift the threshold parameter of the link predictors that we use.

B. Prediction score functions

The link prediction of a collaboration graph is generated by predictor functions for every link in the graph. These functions use the present state of the graph (which includes history from the past, in our case, by using a dynamic graph), and give a prediction score for every possible link to exist in the graph in a future timespan. The higher score represents higher chance for a social link to exist between two nodes. There are numerous models [9] which consider a wide range of possible underlying processes driving the social interactions. Here, we do not aim to debate these models, but to compare their prediction performance, and see, which one fits best to Wikipedia. In the following subsections, we will select five of these models, and give a short overview of them.

Many of the predictors utilize the notion of the *neighborhood* of a node. These are simply the set of nodes adjacent to a given node, which are connected by a social link. The mathematical definition is the following:

$$\Gamma(x) := \{y : x \text{ is socially linked to } y\} \tag{1}$$

1) Common neighbors predictor: The most simple prediction that can be made based on the neighborhood of nodes is the number of common neighbors shared by two nodes. The underlying idea is that the more common neighbors are present, the more chance that the two people will find a common subject, upon which a social link can form between them. The prediction score is defined as follows:

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{2}$$

2) Adamic/Adar: In their paper, A. Adamic and E. Adar proposed, that friendship between two persons can be predicted by measuring their similarity to each other [10]. The similarity is simply measured by the number of shared items, but weighted, such that the unique items (shared only by these two people, and not by others) is more valuable, i.e. gives a stronger prediction, than the item which is shared among many people. Items, in our case, correspond to people, specifically the friends already present at the given time. Therefore, the prediction score given by this predictor is defined as follows:

$$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \tag{3}$$

3) Jaccard's coefficient: This is a similarity metric of sample sets. Generally, it defined as the size of the intersection of two sample sets divided by the size of the union of the sample sets. In the application of link prediction, the samples are the neighborhood of two nodes. From a probabilistic viewpoint, the score is again based on the number of common neighbors, but it is weighted by the probability that a (uniformly) randomly selected neighbor of either nodes is actually a common neighbor of both nodes. The score function is defined as:

$$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{4}$$

4) Preferential attachment: It is based on the growth model of social networks; the basic idea is that a new edge has a probability to be incident on a node is proportional to the current neighborhoods size of that node. In case of collaboration networks, it is suggested that new links form with probabilities proportional to the product of the neighborhood sizes of the two endpoints of a link [8]. Therefore, the score function is defined as:

$$score(x, y) = |\Gamma(x)| \times |\Gamma(y)| \tag{5}$$

5) Weighted common neighbors: We have also added our own predictor, which is an extension of the common neighbors predictor, designed specifically for the dynamic collaboration graph. In order to utilize the present link strength information, beyond whether it's above or below the link threshold, we incorporate the link strengths as weighting factors for the prediction score. The idea is that if social links exist between common neighbors of two nodes, then it can be expected that the probability of a link formation between them is proportional to the strength of present social links to these common neighbors. The prediction score is defined as:

$$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} S(x, z) S(z, y), \tag{6}$$

where $S(x, y)$ denotes the current strength of a link (between endpoints x and y).

C. Binary predictions

Although link predictor functions give integer score values, they are treated as binary predictors, which means that they have either a positive or a negative prediction (i.e. the link will form, or not), and they are compared to binary outcomes (i.e. the link actually formed, or not). This is done by setting threshold parameters for both the predictions and the social links. If the prediction score is above the *prediction threshold*, it is a positive prediction (social link predicted to form), otherwise, it is a negative prediction (social link predicted not to form). If the link strength in the collaboration graph is above the *link threshold*, it is a positive (social) link, otherwise it is a negative link (no social link). The prediction thresholds are different for each prediction model. Since the link strength is time-dependent, the existence of social links is also time-dependent, with the same time resolution as the dynamic graph.

D. Predictor analysis

After setting the link and prediction thresold parameters, the predictor can be applied at a given time step of the dynamic collaboration graph, at any given link. Nonzero prediction score corresponds to an actual prediction, which is classified as positive or negative using the prediction thresold, and compared to the actual future state of the dynamic graph, with a given ΔT time between the present and future. One prediction with the corresponding actual outcome is considered one sample.

The samples are collected over the time period of the first 150 weeks of Wikipedia editing records, for every possible link in each time step. For their analysis, we use statistical tools borrowed from signal detection theory and predictive analytics. Since we have a binary classification of predictions and outcomes, we can use the confusion matrix (Fig. 1) to accumulate the samples. This is a 2×2 matrix, showing the number of samples that have fallen into each of four possible outcomes. From this table, we can derive a number of statistics, defined as follows:

- sensitivity = TP / (TP + FN)

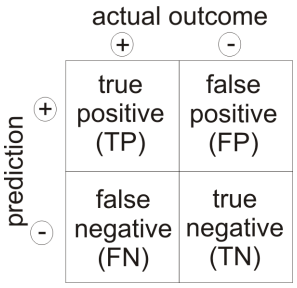

Fig. 1: Confusion matrix of a binary predictor. It contains the number of samples corresponding to each possible outcome.

- specificity = TN / (FP + TN)
- precision = TP / (TP + FP)
- negative prediction value = TN / (TN + FN)
- accuracy = (TP + TN) / (TP + TN + FP + FN)

These quantities can also be defined using conditional probabilities, which give further insight into their meaning:

- sensitivity = $\Pr$(positive prediction | link will form)
- specificity = $\Pr$(negative prediction | link will not form)
- precision = $\Pr$(link will form | positive prediction)
- negative prediction value = $\Pr$(link will not form | negative prediction)
- accuracy = $\Pr$(prediction = outcome)

These is always a tradeoff between sensitivity and specificity, depending on the prediction threshold that we use. The ROC curves (Receiver Operating Characteristic, [11]) show this exactly, by plotting sensitivity against (1−specificity). The advantage of these plots is that they directly visualize the screening capability of the predictor. A random guess prediction would have a point along the diagonal line on this plot, but a perfect predictor would be the point at the top left corner (at coordinates $(0, 1)$), having maximum sensitivity (no false negatives) and having maximum specificity (no false positives). We can plot ROC curves by computing statistics at different prediction threshold parameters, and see which predictor (at which threshold parameter) can get closest to the maximum sensitivity/specificity point.

However, this only tells us the screening capability of a predictor. We also need to know its accuracy, the actual success rate of the predictor. More specifically, we need the precision (also known as positive prediction value) and the negative prediction value, because we can expect a very large number of correct negative predictions, which would significantly influence the accuracy, while we are more interested in positive predictions (social link formations). Precision and accuracy together however are enough for a complete description (besides sensitivity and specificity), we don't need the negative prediction value as a separate third quantity.

We will also use another measure of the prediction quality, the F_1 score. The formula is defined as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}} \tag{7}$$

It values both sensitivity and precision equally, therefore it gives an overall measure of predictor performance.

E. Prediction sampling algorithm

Three predictors, namely the common neighbors, Adamic/Adar, and Jaccard's coefficient predictors need to enumerate all common neighbors of all pairs of nodes, to generate all predictions at a given time step. The naive solution would be to loop over all possible node pairs, and compute neighborhood intersections. If the number of nodes, and the average degree of a node are denoted by N and D, respectively, then the expected run time of this solution would be $\mathcal{O}(N^2D)$, assuming adjacency list storage for the graph, and a hashset for computing the node neighborhood intersection. Since the number of nodes is over three million, this is not feasible.

A better solution is to focus on the common neighbors themselves, and take advantage of the sparseness of the graph. There are many node pairs which do not share any common neighbors, therefore there is no prediction for them at all, so we should not include them in the enumeration. Instead, we enumerate the nodes only once, and look at their neighborhood: any pair of edges incident on a given node, having strength larger than the link threshold, will give a prediction for the link connecting the endpoints of those edges. Therefore, we only need to find all triangles centered on a given node. The complexity of this enumeration is only $\mathcal{O}(ND^2)$, and assuming that the graph is sparse at any given moment, such that $D < \sqrt{N}$, this is better than $\mathcal{O}(N^2)$. The pseudocode for this method is given in Algorithm 1, where $fraction(i, x, y)$ is defined according to the given prediction model; it gives fractional scores based on node x, y, and their common neighbor i. For example, $fraction(i, x, y) = 1$ regardless of parameters for the common neighbors method; $fraction(i, x, y) = 1/Log(Length(L))$ for the Adamic/Adar predictor; $fraction(i, x, y) = 1/|Neighbors(x) \cup Neighbors(y)|$ for the Jaccard's coefficient.

To make the comparisions between predictions and actual future, we do not store the entire dynamic graph of collaborations; it would require too much memory. Instead, we store two snapshots of the graph: one for the "present" time step, and one ΔT time in the future. Both instances are updated simultaenously using the input data of page editings, applying the same update rules at every time step. They are both initialized from the same state (an empty graph), but the future instance is advanced by ΔT time before the predictions and comparisions begin.

Algorithm 1 Score by common neighbors

```
for all node i in graph G do
  L := LIST
  for all node j in neighbors of i do
    if strength(i, j) ≥ linkThreshold then
      add j to L
    end if
  end for
  if Length(L) ≥ 2 then
    for j := 1 to Length(L) do
      for k := j + 1 to Length(L) do
        score(L(j), L(k)) += fraction(i, L(j), L(k))
      end for
    end for
  end if
end for
```

III. Results

A. Static graph properties

As a preliminary analysis, static collaboration graphs were also analyzed. In these graphs all the page editing records are integrated over the entire input data timespan (roughly ten years). There are 3.1 million nodes in these graphs, the total number of distinct editors in the dataset. Links are present between editors if they edited a total of k pages together, where k is a threshold parameter. Figure 2 shows the degree distributions of these graphs. We were interested if we could find a scale-free degree distribution, but we found that nodes with small degrees follow a different scaling than high-degree nodes. This may suggest that the very active editors (high degree nodes) are driven by a different social process than the rest of editors (low degree nodes).

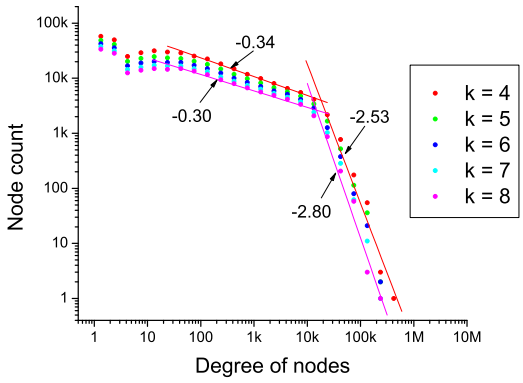

Fig. 2: Histogram of node degrees in the static collaboration graphs, integrated over the time of the entire dataset. The degree distributions were logarithmically binned, and the bins are normalized by their size, such that the histogram is proportional to the original degree distribution. Lines are fitted to different segments of the distributions, the slope of these lines is indicated by the numbers on the figure. Values of k are the minimum number of pages that two editors edited together.

B. Parameter space mapping

For a complete analysis of the selected predictors, we have to consider all input parameters that define the prediction. These are the ΔT time between the present, when the prediction is made, and the time in the future, for which the

prediction is made; the link threshold value, which decides the edge strength, above which the edge is considered as a social link; and prediction threshold of the score, above which it is considered a positive prediction, and below it is a negative prediction. To map this three-dimensional parameter space, we use the following range of parameter values:

- $\Delta T \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ (weeks)
- link threshold $\in \{30, 60, 90, 120, 150\}$

The prediction threshold is different for each predictor, the range must be selected such that the tradeoff between sensitivity and specificity is measurable.

The analysis shows that for all predictors, the link threshold and ΔT parameters have very little influence on the quality of the predictor. We use this to simplify the display of the results. To display the dependence on ΔT and link threshold values, we use contour plots that show the achievable maximum sensitivity (also, maximum specificity), maximum accuracy, and maximum F_1 values. These maxima were found by numerically scanning the range of the prediction threshold values, for the given ΔT and link thresholds. These plots are organized into a table of figures, Fig. 3.

C. Statistical analysis

The statistical behaviour of each predictor is shown concisely on ROC-space plots. For every analysis (for every predictor) two parameters are fixed: $\Delta T = 3$ weeks; link threshold = 90. Then, by running prediction analyses for a range of prediction thresholds (different for each predictor), the measured values of sensitivity, accuracy, precision, and F_1 values are plotted against (1−specificity). These points joined together make continouos curves.

1) Prediction using common neighbors: The measured statistical quantities of the common neighbors predictor is shown in Fig. 4. The following prediction threshold values were used:

$$\text{prediction threshold} \in \{10^{0.1i}\}_{i=0}^{30} \tag{8}$$

Higher values correspond to higher precision and lower sensitivity, but the relation is nonlinear. The curves do not extend beyond specificity value of 0.45, because this corresponds to the prediction threshold (number of common neighbors) $= 1$. This is the minimum possible value, so we cannot get statistics beyond this limit.

The ROC curve shows a good level of sensitivity, but the maximum sensitivity and specificity point (the one closest to the top left corner), which is 83% sensitivity, corresponds to very low precision. On the other hand, we can achieve very high precision, almost 100%, by using a very high threshold, but in this case the sensitivity will be very low (i.e., many false negatives). Overall, the best quality (maximum F_1 value) is achieved at prediction threshold $= 15.8$, where the precision and sensitivity are both 60%.

2) Prediction using Adamic/Adar: The analysis of this predictor is shown in Fig. 5. The range of threshold parameters used:

$$\text{prediction threshold} \in \{10^{0.1i}\}_{i=-20}^{20} \tag{9}$$

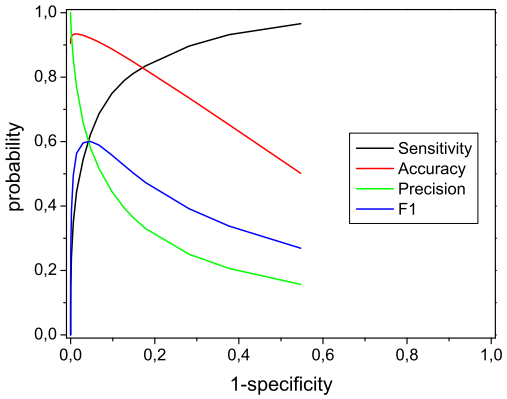

Fig. 4: Statistical behaviour of the common neighbors predictor, shown on the ROC-space.

Higher values again correspond to higher precision and lower sensitivity. In this case, however, the range of thresholds allows to find statistics across the entire ROC-space, so the curves on the figure fill the entire $(0, 1)$ range of specificity.

We can see a somewhat stronger ROC curve, compared to the simple common neighbors predictor. Maximum sensitivity and specificity is at 85%, but again, this corresponds to low precision. The overall best prediction is achieved at prediction threshold $= 3.98$, where both precision and sensitivity are at 62%.

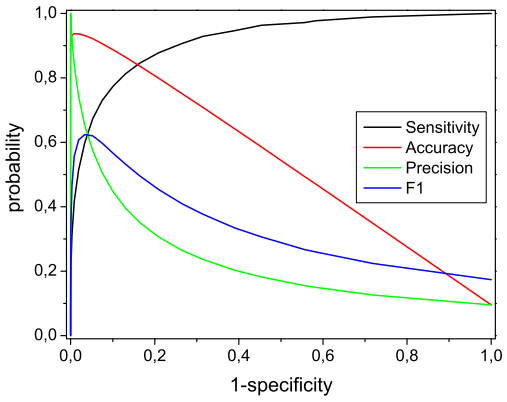

Fig. 5: Statistical behaviour of the Adamic/Adar predictor, shown on the ROC-space.

3) Prediction using Jaccard's coefficient: The ROC-space plot is shown in Fig. 6 for this predictor. The range of threshold parameters used to generate the plot:

$$\text{prediction threshold} \in \{10^{0.1i}\}_{i=-40}^{0} \tag{10}$$

Note, the maximum possible value of this coefficient is 1, which corresponds to the situation where the two nodes only

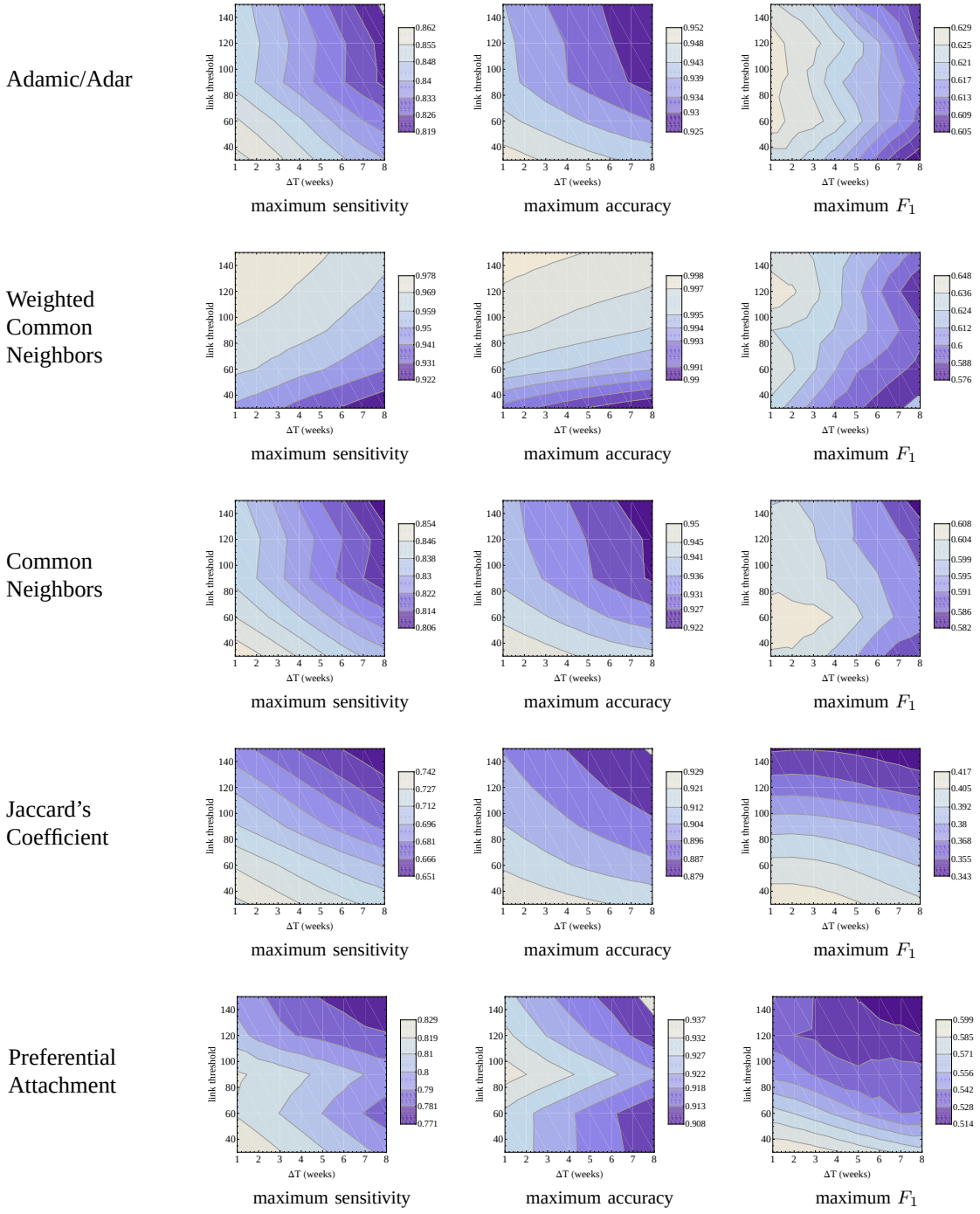

Fig. 3: Maximum achievable sensitivity, accuracy, and F_1 values, as a function of ΔT time (between prediction and actual outcome), and link threshold parameters, for each predictor.

have common neighbors.

The statistics show that this predictor has much worse characteristics than previous ones. Maximum sensitivity and specificity is only 70%. The precision and accuracy values do not reach 100% as the prediction threshold increases, the maximum possible accuracy is 90% at threshold $= 0.50$, and maximum precision is 48%, found at the same threshold value. The best overall performance is found at threshold $= 0.25$, where precision and sensitivity are 38%.

4) Prediction using Preferential attachment: The analysis of this predictor is shown in Fig. 7. The range of threshold parameters:

$$\text{prediction threshold} \in \{10^{0.25i}\}_{i=0}^{16} \tag{11}$$

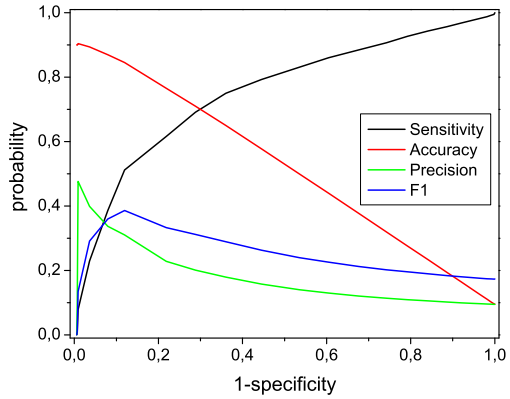

Fig. 6: Statistical behaviour of the Jaccard's coefficient predictor, shown on the ROC-space.

Note, that this predictor's lowest possible value is 1, when both nodes have only one neighbor. Higher values correspond to higher precision and lower sensitivity. The curves do not extend beyond specificity of 0.1, becuase the threshold value $= 1$ corresponds to this point, and it can not be smaller.

The ROC-curve is somewhat better than for the Jaccard's coefficient, but it's worse than the common neighbors, the maximum sensitivity (and specificity) is 81%, achieved at prediction threshold $= 30$. Precision can reach nearly 100%, but like in case of other predictors, it would result in very low sensitivity. The overall best prediction is achieved at prediction threshold $= 100$, where the sensitivity and precision are both 54%.

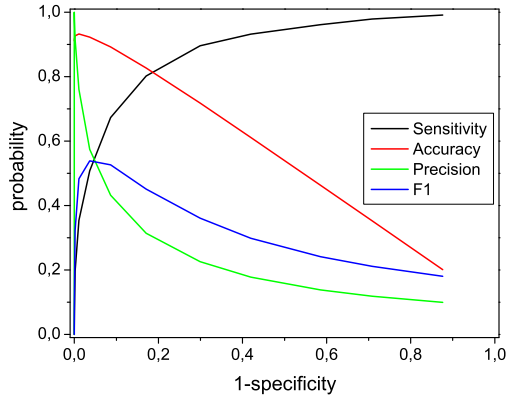

Fig. 7: Statistical behaviour of the preferential attachment predictor, shown on the ROC-space.

5) Prediction using weighted common neighbors: Finally, the analysis of our proposed predictor is shown in Fig. 8. The following range of threshold parameters was used:

$$\text{prediction threshold} \in \{10^{0.25i}\}_{i=4}^{28} \tag{12}$$

Note, that the lowest possible prediction value now depends on the link threshold: prediction value $\geq$ (link threshold)^2. It was observed, that some links manage to gain strength in the order of thousands, so the highest values of predictions usually range in the millions.

The analysis shows, that this method has superior screening capability. Its maximum sensitivity (with maximum specificity) exceeds all other predictors: 97%, when prediction threshold $= 150000$. It also has a very good overall performance: 61% is the maximal precision and sensitivity, at prediction threshold $= 1.7 \times 10^6$.

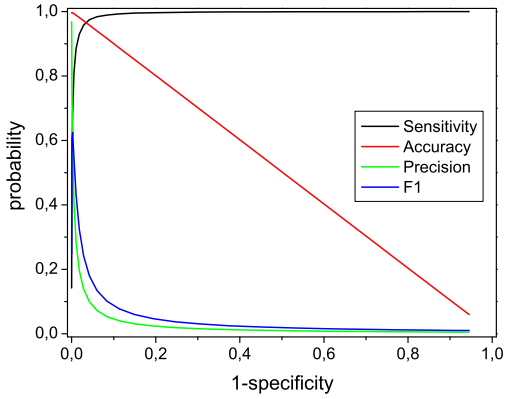

Fig. 8: Statistical behaviour of the weighted common neighbors predictor, shown on the ROC-space.

D. Comparision of predictors

There are three statistical aspects for comparing the predictors. We can strive for maximum accuracy, if we want to be most correct in our predictions for the future. Alternatively, we may look for the best screening method, which is able to identify most of positive and negative link formations correctly for the future. In other words, we can strive for maximum sensitivity and specificity. The third option is the golden mean of the two previous goals: If we need a predictor that is both highly sensitive and highly precise, then we look for the maximum achievable F_1 ratio. According to these cases, Figs. 9, 10 and 11 compare the examined predictors to each other.

For maximum sensitivity, the weighted common neighbors is clearly the best method. It also means that the actual strength of existing social links are indeed very important in the similarity computation between nodes, and this information is more precise at every given time step, if information about the past is included.

In case of maximum accuracy, we must be careful to correcly interpret Fig. 10. The weighted common neighbors is theoretically capable of achieving nearly 100% accuracy,

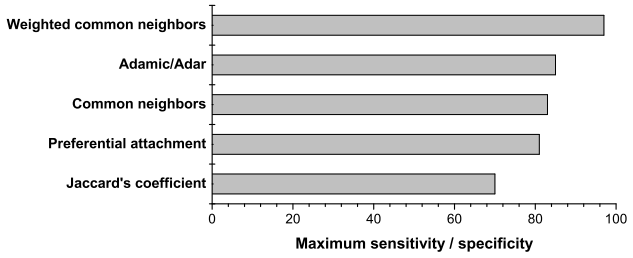

Fig. 9: The maximum achievable sensitivity and specificity values with the predictors.

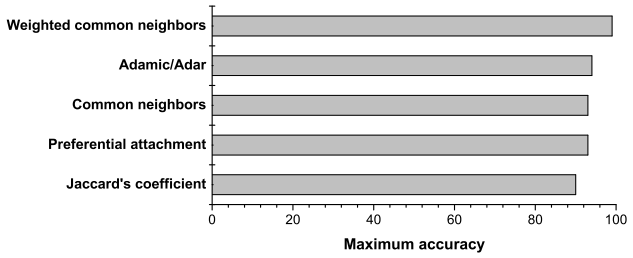

Fig. 10: The maximum achievable prediction accuracy with the predictors.

but this is because it gives a very large number of correct negative predictions (inherently, because of the sparseness of the network), while the rate of correct positive predictions is diminished. Note, however, that the next best is the Adamic/Adar method, which has sufficiently high precision at its highest accuracy value.

When considering maximum overall performance, the best F_1 score is attained by the Adamic/Adar method. However, it is notable, that the weighted common neighbors is also very close, and in fact it can achieve even higher scores, when the link threshold is higher, and ΔT is smaller, see Fig. 3.

IV. Conclusion

If we consider that all predictors have predicting capability well beyond a random guess, it is clear, that there is a strong social process between the editors of Wikipedia. This is, however, a complex process, which involves many human factors. The best predictors shown here can capture most of these factors by using careful assumptions about the strength of social links between editors, derived from noting but collaborations on edited pages.

When looking for a predictor, one must always be clear about the goal that he wishes to achieve. We have seen that maximum sensitivity needs different parameters than maximum accuracy. Computation of these statistical properties revealed that the best screening method is the weighted common neighbors, and the most accurate predictor is the

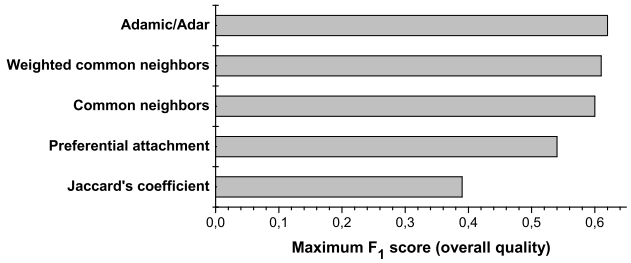

Fig. 11: The maximum achievable overall quality (F_1 score) with the predictors.

Adamic/Adar predictor, among the examined methods in this paper. By overall quality, Adamic/Adar gives the most precise prediction with the highest sensitivity.

Acknowledgment

The author would like to thank professor Malik Magdon-Ismail (Computer Science department, Rensselaer Polytechnic Institute) for providing the Wikipedia dataset, and valuable lectures on computational analysis of social processes.

References

[1] A.-L. Barabási, R. Albert, *Emergence of scaling in random networks*, Science, 286(5439), 509512, 1999.
[2] M. Granovetter, *The Strength of Weak Ties*, American Journal of Sociology, 78(6), 1360–1380, 1973.
[3] E. M. Jin, M. Girvan, M. E. J. Newman, *The structure of growing social networks*, Physical Review Letters E, 64(046132), 2001.
[4] T. Tylenda, R. Angelova, S. Bedathur, *Towards Time-aware Link Prediction in Evolving Social Networks*, Proceedings of the 3rd Workshop on Social Network Mining and Analysis, 2009.
[5] Z. Huang, D. Lin, *The Time-Series Link Prediction Problem with Applications in Communication Surveillance*, INFORMS Journal on Computing, 2008.
[6] H. H. Song, T. W. Cho, V. Dave, Y. Zhang, L. Qiu, *Scalable Proximity Estimation and Link Prediction in Online Social Networks*, Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, 2009.
[7] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabási, *Human Mobility, Social Ties, and Link Prediction*, In Proceedings of the 17th ACM SIGKDD intl. conf. on Knowledge discovery and data mining, 2011.
[8] A.-L. Barabasi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, *Evolution of the social network of scientific collaboration*, Physica A, 311(3), 590–614, 2002.
[9] D. Liben-Nowell, J. Kleinberg, *The link-prediction problem for social networks*, J. Am. Soc. Inf. Sci., 58(7), 1019–1031, 2007.
[10] L. A. Adamic, E. Adar, *Friends and neighbors on the web*, Social Networks, 25(3), 211–230, 2003.
[11] K.H. Zou, A.J. O'Malley, L. Mauri, *Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models*, Circulation, 115(5), 654-657, 2007.