JKRW Link Prediction

A New Ensemble Technique Based on Merging Other Known Techniques in The Social Network Analysis

https://doi.org/10.3991/ijim.v15i12.22831

Aya Taleb, Rizik Al-Sayyed, Hamed Al-Bdour
The University of Jordan, Amman, Jordan
aqrabawi.aya@gmail.com

Abstract—In this research, a new technique to improve the accuracy of the link prediction for most of the networks is proposed; it is based on the prediction ensemble approach using the voting merging technique. The new proposed ensemble called Jaccard, Katz, and Random models Wrapper (JKRW), it scales up the prediction accuracy and provides better predictions for different sizes of populations including small, medium, and large data. The proposed model has been tested and evaluated based on the area under curve (AUC) and accuracy (ACC) measures. These measures applied to the other models used in this study that has been built based on the Jaccard Coefficient, Katz, Adamic/Adar, and Preferential attachment. Results from applying the evaluation matrices verify the improvement of JKRW effectiveness and stability in comparison to the other tested models. The results from applying the Wilcoxon signed-rank method (one of the non-parametric paired tests) indicate that JKRW has significant differences compared to the other models in the different populations at **0.95** confident interval.

Keywords—Link Prediction, Network Analysis, Ensemble, Machine Learning, Graph Analysis, Voting Techniques

1 Introduction and Background

1.1 Introduction

Links between the different networks are now the most important factor to predict the next generation of any of the future transactions, common interests, future friendships, communications, and many other fields. However, the most interesting part is the question that has raised, how to properly use the magic of these network links in predicting future links in the most optimal, fast, and accurate way possible to start building facts and businesses based on these predictions. This question is one of the reasons for doing this research study. Many studies and proposals have been talked about networks link prediction and how to utilize and devote it to be used in many businesses and important fields in different life areas. Other studies that were looked

up, have been focused on how to produce the most effective technique and apply it to the predictions aiming to have more accurate and consistent results for better solutions.

Different link prediction techniques and algorithms have been introduced to analyze the current networks and try to predict the possibility of future links and relations between the disconnected nodes in the different networks. These techniques have been used as well to predict the possibility of disconnecting nodes that are connected by a relation in any network [1].

Link prediction is used for many purposes. One of its common usage is in social networks for suggesting new friends. It is also used in more sensitive fields such as the security field to predict future links between the different terrorist groups and increase security measures. Also, it is used in the medical field to predict the relations between any symptoms and diseases for great efficiency and responsiveness. It is moreover used in the biology and bioinformatics fields for greater advancements in the field. For the importance of having more accurate techniques in predicting the links and since the computational tools have become less important than the efficiency; this due to the more powerful computation machines become available that are capable of doing wonders with the ability to benefit from the power of the clustering and threading. So this study aims at introducing a new link prediction technique that can scale up the accuracy and provide better predictions and expectations than the traditional ones regardless of the time factor.

There are great efforts put in improving the prediction techniques and to provide better link predictions for the problems faced so far. However, these solutions and improvements do not work with every type or size of any new network. It solves the problem it has solely designed for and does it with improved efficiency and effectiveness. Each time the researchers need to build several models to compare their performances and choose the one with better prediction results which is heavy and time consuming. On the other hand, there are data-related solutions that can be useful for one dataset and poor for another dataset with the same type but different sizes. Therefore, there is a need to come up with a solution that is working and improved for the majority of data, so that the need for many solutions to solve one problem can be avoided and replaced by having a single solution for most datasets.

1.2 Background

The Internet world develops rapidly in social and web networks. A vast amount of open data and nodes produced by smart systems bring new challenges to cope and work on them daily. Graphs(G) are sets of objects formed from the internet data that have direct or indirect connections between them. Each object in these graphs called a vertex, while the connections linking these vertices called edges. On the other hand, networks are one of the graphs structured representations, vertices called nodes in the networks and the links between them could be directional (asymmetric) or bi-directional (symmetric) [2]. The main characteristic of these networks is having it as a real collection of nodes and links that are dramatically increasing in size and shape over time to produce dynamic and complex networks [3].

Social network prediction is one of the most common and important challenges that the world faces. The ability to predict any new connection between nodes (it can be users, products, or any other item in the network) is something every researcher tries to experiment and provides new techniques or algorithms that can analyze and predict these links in an accurate and improved method. Although these researches produce valuable techniques and algorithms to solve the link prediction problems, these developed algorithms are not enough to fit all kinds of networks and provide a stable accuracy that is required [4].

There are lots of fields like the biology, sociology, diseases, and web-based systems that can use the networks to describe them [5]. In such systems the graph nodes represent individuals, and the edges are the connections and interactions between these nodes. The social networks (SN) formed in a dynamic structure that is forming itself over time with the huge nodes' by adding and deleting different types of links between them. Many papers and research are focusing on this topic of the dynamic network evolution like [6], [7] and, [8].

The complex structure and categories of the shaped networks are a common field for the researchers to study and make some substantial improvements to it. Euler is one of the first researchers who has studied the network bridges case and has come up with great contributions that have been used by almost every researcher afterward in the well-known problem of the 'Seven Bridges' theory of Königsberg [9]. The random network's evolution (ER-model) has been presented by Erdos-Renyi [10]. This leads to the huge data soared up and more problems surged to the complex network processing and the hybrid and heterogeneous graphs [11]. Other techniques were to improve new better solutions based on combined best algorithms as what Salihoun did in their research [12].

The link prediction is a technique to predict if there would be any possibility for a new connection between any random nodes, based on some history and features related to the nodes, to calculate the possibilities of this connection. It considered a huge addition to the social communities as it is not just filling the gaps because of the missed data; it is also used to predict any future link that might be added or removed from the current network.

In the social media network link prediction, the link prediction techniques can help the users to find the friends that are most likely similar to them (i.e. mutual friends on Facebook and friends suggestion on Instagram and any other social networking site). These generated predictions in the social media fields increase the trust between the users and the social network (SN) and increase the loyalty of the social network website providers [13].

Link prediction algorithms used to predict an existence of transaction between different node. A possibility can be calculated based on a specific rank to indicate the probability of having this link or transaction in the future. The higher the prediction rank is the higher probability to have this transaction [14].

The link prediction is being used in three main forms and purposes: (1) predicting a new link between two nodes/graphs, (2) predicting a removal of an existing link between nodes or graphs, and (3) replacing an existing link with another one with an opposite direction.

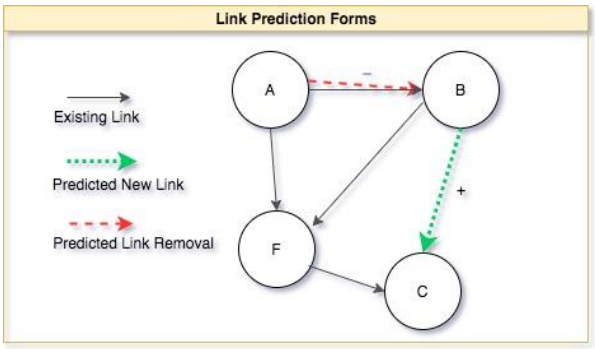

Fig. 1. Three main possible forms of the link prediction between different nodes in any graph

1.3 Datasets details and comparison

The chosen datasets are different from each other in three main factors. These factors are the data source type (social networks and e-commerce ones), population (small, medium, and large), and the number of connected components in each one. Table 1 presents a comparison between these data.

Table 1. Datasets Detailed Comparison

	Facebook	**Amazon**	**Last.FM**
Type	Social networks, Undirected Graph	eCommerce community, Undirected Graph	Online Communities, Undirected Graph
Source	Data from a survey using Facebook app.	Data collected by crawling Amazon website.	Crawled on 2010 by Megan Kearl
Graph Length	3,959	334,863	108,493
Total Nodes	3,959	334,863	108,493
Total Edges	84,243	925,872	5,115,300
Connected Components	13	1	34
Avg. Clustering	0.5437	0.3967	0.0727
Avg. Centrality	2022.5128	276768.5657	597644.8274

1.4 Data cleaning and processing

The data used in this research has been gathered originally in three different files: the first file is for the nodes, the second one is for the edges between nodes, and the last file is for the features per node. The pre-processing step in link prediction problems is usually processed by eliminating and removing the blank nodes which are the nodes without any linked node. Also, to remove all unnecessary features from the files. In this research, the most and important part of the data is the graph nodes and links. Thus, the purpose of this step is to remove the features dependencies and to introduce a new binary column indicating the presence of links between nodes; if any.

In other words, the new column has been introduced to all datasets and then some unconnected entries have been added for nodes without any edges between them, to have the desired trained model on semi-balanced data with better accuracy and more reliable results.

This specific step of processing is important, which adds some negative samples to the final datasets. So, in the end, each dataset has one giant graph where each node is connected to at least one node in the training sets and disconnected from many other nodes. After the data processing, training data which is 70% of each graph has been produced and made ready for training the models.

There is a limitation in Last.FM dataset population regarding its size, because of its large population it was not fit in Python plotting memory to draw it.

1.5 Model development

Since datasets are ready to be used in any research, the model design and preparation can be started based on it. Three graphs were generated based on the population size and type of the source it gathered from, these datasets are clean and ready to conduct and train the several models based on it. On the other hand, there should be clear experimentation factors and variables needed to be considered and defined before starting any random trials. This thesis model development starts with choosing the main set of algorithms that are supposed to have some common characteristics and different working schemes. Those algorithms have been used in establishing several models called in this research as stand-alone models. A stand-alone model is a model that has been built based on only one machine learning algorithm for training and testing it with multiple tuning parameters and different types of input datasets for training.

The way of picking the set of algorithms depends on some factors. Since this research focus is on the accuracy and efficiency regardless of the training offline time (the training phase of the model is conducted in an offline server's mode), then selecting the algorithms to be used in building the final proposed ensemble would be based on some factors that have less dependency on time. Below are the main factors used to select the algorithms included in the proposed ensemble classifier:

- Algorithms that are well known and mostly used in the link prediction models.
- Algorithms that are different from each other in the way of working and the inner prediction techniques.
- Algorithms with high performance and mostly better classifiers in generating any prediction for only a small population dataset.
- Algorithms with high performance in generating the predictions for a small to medium population dataset.
- Algorithms with improved performance and well known in generating predictions for the large and dynamic population sets.
- Trial and error validation. This trick can be used after selecting the set of initial algorithms to validate that the selected set is mostly the best combination for this research.

Considering the above factors and having the described machine learning algorithms and its usage in the link prediction, then the algorithms picked to build and train this research models are: Jaccard Coefficient, Katz, Preferential Attachment (PA), Adamic/Adar (AA)

After choosing the main algorithm set, the trial-and-error validation step has been conducted to validate and ensure the fit for the chosen inputs as part of the proposed ensemble. Different combinations of the above algorithms have been applied to the small dataset (Facebook) to validate its ability to work together. As a result, the best combination was to have the above four algorithms together and build different ensemble models based on them. On the other hand, having these four algorithms together in building the proposed model would ensure that the different types of link prediction algorithms have been covered by the proposed ensemble. Thus, supporting the main objective of this research which is to have a classifier that would work for any type of data and any size of the training set and then produce the best results in terms of accuracy.

The proposed ensemble is called JKRW, since it is an ensemble that is like a wrapper for Jaccard Coefficient, Katz, and other random algorithms. JKRW is generated in two ways, the first one is generated based on the voting techniques that have been described in the literature review section, and the other form is generated based on the averaging technique.

Python is one of the most commonly used languages by data scientists and machine learning communities. It's flexibility, easy to use functionalities, and wide variety of reusable open-source libraries are the main characteristics to choose this language. Therefore, the need for a powerful and flexible language to conduct the model's training from small to large datasets is the main reason to choose Python as the main language to build and evaluate the proposed models.

1.6 Evaluation and assessment

As a result of having four different algorithms set (as defined in the previous section), the four different stand-alone models have been generated for each dataset in addition to the new two proposed ensemble models (JKRW) that been produced based on different ensemble techniques. This has led to the need for a mechanism to evaluate these models in terms of accuracy and the area under curve.

There are several ways to measure the ACC (accuracy) of a model, the simplest one of them is the confusion matrix that has been described in the previous chapter. To calculate the accuracy of any model in Python, a simple library can be used from the Sklearn metrics open-source libraries.

Another result from the models to measure is the AUC (area under curve), to calculate this measure values using Python Sklearn matrix, the model should be first trained and tested using the test data, then the evaluation methods can be applied to calculate the final results (which are presented in the experiment results section). These metrics of AUC and ACC are commonly used in the evaluation of the link prediction problems. After the model building and evaluation phases, the main results

from this research have to be evaluated by comparing the results from the different models.

There are two ways to compare these results. The first traditional one is by having that comparison manually using the resulted values from the described evaluation matrices. The second approach is more scientific method to compare the whole models' accuracy using the statistical significance tests and then apply the AUC after having confidence in the differences between models.

2 Experiment Details

2.1 Experimental study

The main focus of this research is to build stand-alone models for the link prediction problem for each set of data provided. Then, build the ensemble proposed solution by merging the same stand-alone generated models based on some ensemble merging techniques. The results analysis has been conducted to prove that the proposed solution has better improvement and higher usage regardless of the data or network size.

Both link prediction and ensemble approaches and techniques have been studied carefully since these fields were new in the research world and therefore, must be examined properly. Also, related works and research have been studied to look into the exact gap and how other researchers tried to overcome it in the link prediction world. With the lack of features and huge connected networks, link prediction became more and more difficult but, at the same time, an important type of prediction. Another factor for consideration was the e-commerce and social networks problems. And how these fields have become the leading problems and concerns in the large data evolution and smart system’s needs. Data used in this experiment has been collected from an open source, solely for research purposes. Some other types of well-defined data sets and networks have been requested from other types of private sources, but it got either rejected or limited with the number of nodes provided. This fact moved the study effort and focus on the open free data sources based on the required needs.

2.2 Experiments settings

To conduct this research, a tool needed to build machine learning models in all environments. Python has been chosen as a language for its flexibility, scalability, and the ability to handle any number of clusters and threads in parallel. Machine learning libraries and ready to use algorithms have been used from Python like pandas, networkx, numpy, sklearn, matplotlib, and other Python packages related to the run-time and distribution improvements while executing the program and building the models.

A sample network data has been used which contained a small number of nodes and edges generated randomly using Network-x graph methods to evaluate the proposal before digging deep into it. Then, the models have been trained and built based on that simple dataset to prove the ability to continue with the research. The conclu-

sion of these steps has nothing to do with the performance comparison but instead, it served as proof that this study worth the value and can fill the link prediction accuracy gap that it is supposed to fill.

2.3 Experiments execution and limitations

The execution has been done iteratively for each algorithm. The first group of models have been built based on Jaccard Coefficient algorithm for each one of the three datasets. Run time values were clear to notice during the first few runs, especially for the large data sets. The second group of models have been built based on Katz algorithm while the third group used the Preferential Attachment algorithm in training and building the models. The final standalone models have been built based on Adamic/Adar technique. After the first round of models' setups and building, twelve weak/standalone models have been built and ready for the next steps.

The second phase has been conducted to build the proposed ensembles for each dataset. The first group of ensembles called JKRW-Avg have been built based on the above four models and by combining them based on the simple averaging technique. In this technique, the model has been used to generate predictions based on the calculated average prediction results from each one of the input models used for building the ensemble. Another group of the proposed ensembles called JKRW-Voting have been built based on the above standalone models, combining them using the Voting technique. This technique depends on having the final prediction based on the majority of predictions from the whole models. Thus, it usually gives the most accurate and precise results than any single prediction model.

After all rounds of models building, it has ended up with eighteen models on all of the data sources.

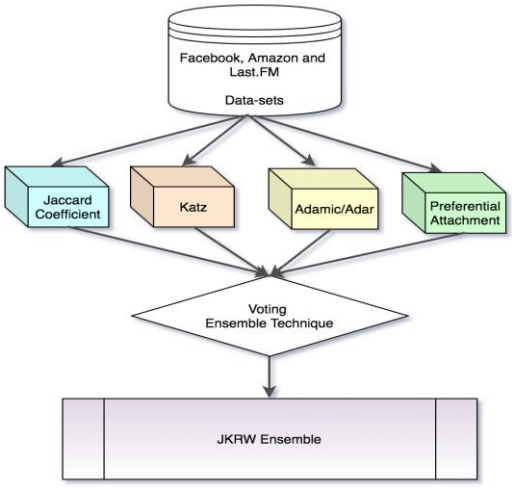

Fig. 2. JKRW Ensemble model in the heterogeneous ensemble's representation

3 Results Discussion and Evaluation

3.1 Evaluation analysis

To evaluate the experimental models, the single models based on link prediction algorithms have been evaluated firstly. Then, the JKRW ensemble models have been evaluated. Then a proper comparison has been conducted based on these evaluation results.

Two ways and techniques of evaluation have been established for each model from the eighteen resulted models. The first one was the traditional evaluation technique via the confusion matrix and accuracy numbers. Both false positives and false negatives were not the highest important factors in the evaluation process. Since the main objective of this research was to enhance the prediction accuracy in the data of the different network.

The other way of the evaluation was the area under curve. This technique is widely used in the link prediction-based models to see and understand the coverage of the models' predictions compared to each other. It has also been used to find the models' ability to distinguish between the prediction output values in the positive values space at different thresholds than the accuracy. On the other hand, the AUC matters in this evaluation since there was a need to ignore any over-fitting problems because of the binary problems got in the link prediction areas. And then to closely look into the output based on the true positive rate (TPR) and false positive rate (FPR) for every single model. The data has been split into training and testing sets for each one of the used datasets. 70% of the data has been used for training the models, and 30% for testing them. The time factor has been calculated as well since some of the models took a long time and consumed memory to train the models. Some other models used way more time, so it has been clustered and trained on different machines with higher memory and capabilities.

3.2 Results analysis and comparisons

As per the evaluation methods used for each model. Many factors have been looked into while producing the results. The accuracy evaluation has been provided for the different datasets. These results have led to start thinking about some links between the accuracy and population size and test size compared to the ensemble results for each population case.

The other two factors that have been looked into when pursuing the evaluation of results and comparing it are the area under curve (AUC) and the time taken for building the models. Since the time was not important in this research scope of improvement, then it has been presented in factor way considering the minimum time with a special formula. The minimum time taken was around twenty minutes, so T representing 20 minutes of training the model.

Looking at the accuracy and area under curve results described in this research, the results have presented a different behavior for each stand-alone model based on the

dataset population size and type. The main reason for these high differences in numbers and behaviors returns to the population size and type used to train and test each stand-alone and ensemble model. On the other hand, having a quick scan to the ensembles JKRW models, the numbers were logical in all cases compared to the stand-alone ones. Having this different range of accuracy and area under curve numbers, a deeper analysis was required on each population type and size of the model groups.

Following the analysis of the results based on the evaluation numbers conducted in this research and grouped by population type to have better results understanding and answers to this research questions:

Small population dataset (Social media random connections) - Facebook Results: Looking at the accuracy and the area under curve results for the four stand-alone. The overall accuracy for this dataset for all algorithms seemed to be low compared to the area under curve for the same dataset and models. The reason behind having this difference was related to the data itself and how these different measures work. Since the AUC has been calculated based on the true positive rate (TPR) and false positive rate (FPR) in the sample data using different threshold values, while the overall accuracy has been calculated using a static threshold and based on all different measures from the confusion matrix (TP, TN, FP, FN). Thus, having this low accuracy in all of the estimators for the small population of Facebook and better AUC numbers for the same estimators means that the data wasn't balanced enough and have a bias to some of the classes.

Table 2. The accuracy, area under curve, and time results for the models trained using Facebook data

Dataset	**Facebook (small) dataset network**					
Model	***Jaccard***	***Katz***	***Pref. Attach***	***A.A***	***JKRW (Voting)***	***JKRW (AVG)***
Accuracy	0.3676	0.3106	0.2965	0.2529	0.3766	0.2094
AUC	0.7529	0.6209	0.5961	0.4949	0.7799	0.789
Time (T Factor)	T	0.5T	0.5T	T	3.5T	3T

On the other hand, the JKRW ensembles in both ACC and AUC have different results based on the ensemble technique used. JKRW-Voting was doing better in terms of ACC than the JKRW-Averaging and better than all other different stand-alone models, while the JKRW-Averaging has lower results than other estimators in the accuracy matter which means that it got the weakest results from the different models shaped it. AUC results for the proposed JKRW ensembles have presented that both ensembles were able to produce better prediction and AUC numbers than all other stand-alone models. However, JKRW-Averaging was doing slightly better with having better AUC value than the JKRW-Voting in the context of a small population for random social media links dataset.

Medium population dataset (e-commerce strong nodes connections) - Amazon Results: Amazon dataset seemed to be the best sample that got logical and close measuring results after applying both ACC and AUC matrices. The reason behind having such close results in the different measures while they have different evaluation techniques return to the fact that Amazon dataset for training and tests population

size was enough to generate better predictions than the small population dataset, another reason related to the type of the graph used from this dataset, which has been pulled from a complete non-randomized data source from the e-commerce field. The Accuracy results, showed that the proposed ensemble JKRW-Voting got better accuracy among other stand-alone and ensemble models. On the other hand, the other ensemble JKRW-Averaging was not doing well (the same as in the small population results). Another conclusion from these results based on ACC, that the Pref Attachment model was doing well and close to the JKRW-Voting ensemble with a slightly lower overall ACC number.

AUC results have been presented as well for the medium population dataset from Amazon. A similar analysis for ACC found here, the proposed JKRW-Voting ensemble was doing better than the other models while the other proposed JKRW-Averaging ensemble has produced poor results compared to all other stand-alone models.

Table 3. The accuracy, area under curve, and time results for the models trained using Amazon data

Dataset	**Amazon (Medium) dataset network**					
Model	***Jaccard***	***Katz***	***Pref. Attach***	***A.A***	***JKRW (Voting)***	***JKRW (AVG)***
Accuracy	0.2633	0.4051	0.6103	0.323	0.699	0.4176
AUC	0.3039	0.6699	0.2489	0.6605	0.6876	0.4683
Time (T Factor)	2T	T	T	T	4T	4T

Large to X-large population dataset (Online graph-based website) - Last.FM Results: These results produced by conducting the experimentation on the x-large dataset of Last.FM dataset, taken the most time of this research experimentation time. Training and building the whole models using this dataset took more than triple the time used for building the models based on Amazon(medium) and Facebook(small). Thus, other computers with better processing and memory were used to run the experimentation, evaluate the models, and getting the results. Since Last.FM data has been collected from a huge graph from the source, and after applying the pre-processing step to it, it became balanced and ready to start experimenting based on it. Results for the ACC measure, showed a close similarity between the proposed JKRW-Voting and the Preferential Attachment based model. However, the JKRW-Voting ensemble still has better ACC results among all other models including the other proposed JKRW-Averaging ensemble. The proposed JKRW-Voting ensemble was doing better than other models in the AUC results; it was even better than the other proposed ensemble JKRW-Averaging. Another notice from the AUC results that JKRW-ensemble was somehow close to the Katz AUC results (with a better measure for JKRW-Voting), which means that JKRW-Voting for the large data used in this experiment was almost close to the better algorithm results. Thus, these results have been used as evidence that the voting ensembles were used generally to choose the better estimator from a different set of algorithms regardless of the population size and type used in the model training and testing, which is the purpose of this study.

Table 4. The Accuracy, area under curve, and time results for the models trained using Last.FM data

Dataset	**Last.FM (X-Large) dataset network**					
Model	***Jaccard***	***Katz***	***Pref. Attach***	***A. A***	***JKRW (Voting)***	***JKRW (AVG)***
Accuracy	0.23017	0.3081	0.5812	0.4921	0.5894	0.4395
AUC	0.4719	0.6156	0.4914	0.4928	0.6329	0.5918
Time (T Factor)	5T	3T	3T	5T	11T	11T

As final results from the above analysis for the ACC and AUC measures results, there were noticeable better results for the proposed ensemble JKRW-Voting among other models of stand-alone generated ones and ensemble proposed ones based on the averaging technique.

The time factor was bigger for all results from this research experimentation, but that seemed to be logical since the whole experimentation was not aiming to reduce the time complexity factor for the result ensembles.

Table 5. Accuracy overall results for the different tested models over different networks

Network	**Algorithm**					
	Jaccard	***Katz***	***Pref. Attach***	***A.A***	***JKRW (Voting)***	***JKRW (Avg)***
Facebook	0.3676	0.3106	0.2965	0.2529	0.3766	0.2094
Amazon	0.2633	0.4051	0.6103	0.323	0.699	0.4176
Last.FM	0.3201	0.3081	0.5812	0.4921	0.5894	0.4395

Table 6. AUC overall results for the different tested models over networks

Network	**Algorithm**					
	Jaccard	***Katz***	***Pref. Attach***	***A.A***	***JKRW (Voting)***	***JKRW (Avg)***
Facebook	0.7529	0.6209	0.5962	0.4949	0.7799	0.789
Amazon	0.3039	0.6699	0.2489	0.6605	0.6876	0.4683
Last.FM	0.4719	0.6156	0.4914	0.4928	0.6329	0.5918

3.3 Models comparison

Comparing different machine learning models based on the ACC results analysis and AUC values was not enough. In the machine learning models, to be able to choose between different models or samples, some statistical methods should be applied to be confident of any efficiency enhancement via differences comparisons. These statistical methods called statistical significance tests. The way to determine the best model among different models depends on the skills and factors needed from this model to be achieved. However, trusting the resulted evaluations estimated from the models based on some test sets was a lack of confidence and trust for all generated models.

In this research and to be confident that the proposed ensemble proved to be the winner and can be used in further real applications with trust, then some statistical significance tests have been applied to the final generated models.

One of the problems faced while choosing from the statistical significance tests was the lack of normal distribution in the samples. Another problem was having different kinds of samples in terms of mean and size. Thus, the traditional t-test (student statistical test) has been replaced with one of the non-parametric paired tests [15].

From the non-parametric paired tests, the Wilcoxon signed-rank test has been used as an indicator for the distributions of the given samples. This study paper presenting the results from applying the Wilcoxon statistics on each pair of models to compare the proposed JKRW with each model in each space of data. Alpha value (the significance confidence attribute), its' default value has been used in this research which was 5% (0.005), that means any p-value more than alpha was not considered to be useful confidence in the improvement's comparisons between the tested models, while the less p-value provides the proposed JKRW ensemble with better confidence in its hypothesis of improvements.

The Wilcoxon signed-rank results are discussed in this research. There were twelve comparisons conducted. Two out of these comparisons failed to prove the required confidence in the proposed model which are JKRW Vs Jaccard using Facebook data and JKRW Vs Preferential Attachment in Last.FM data space. Therefore, there were ten out of twelve comparisons with a p-value less than alpha (0.05) which are the majority of the conducted comparisons. Thus, as a conclusion, the proposed ensemble JKRW have a significant difference compared to the other stand-alone models in the different population sizes and types at 0.95 confident interval.

Table 7. The Significance comparison of the proposed ensemble scores with the other studied models using the Wilcoxon signed-rank test

	Small Data (Facebook)	**Medium Data (Amazon)**	**Large Data (Last.fm)**
JKRW Vs Jaccard	0.7083	0.0046*	0.0092*
JKRW Vs Katz	0.0455*	0.011*	0.0013*
JKRW Vs Pref. Attach	0.0447*	0.0412*	0.0705
JKRW Vs Adamic/Adar	0.0026*	0.0081*	0.0461*

4 Conclusion and Future Work

4.1 Conclusion

The proposed JKRW ensemble approach for the link prediction problems has scaled up the accuracy and efficiency of the social and e-commerce link predictions.

The experiment results and evaluations were evident of the improvement in the accuracy and coverage for this proposed model compared to the other models used in the link prediction-based models. It proved these enhancements and increased the level of model trust in all kinds and sizes of data populations. Furthermore, this research has scrutinized two types of proposed ensembles based on the majority voting and the simple averaging combination techniques.

This research then proved that there is a slightly greater preference in terms of better results and higher accuracy in the JKRW-Voting over the JKRW-Averaging.

Finally, based on the results and evaluations of the hands-on experiments, JKRW-Voting has been verified to have better and more accurate predictions in the link prediction problems than the existing most common stand-alone algorithm-based models.

4.2 Future work

A couple of areas need more research work to be conducted and can be studied deeply and, in more depth, and scope. First, other new versions of this ensemble could be produced by adding one or more algorithms, or by adding different ensemble merging techniques.

The other area of enhancement is having further investigations on how to improve the time performance of the JKRW without the need to drop any of the required processes. The security and privacy of the link prediction techniques could be a focus for further research and studies. Thus, JKRW could be studied from a different aspect of improvements related to security and privacy matters while producing the predictions. Finally, Implementing and applying JKRW prediction model in an actual application and start researching and studying the ability to use continuous feedback to continue improving this model and making it more reliable.

5 References

[1] Gupta, S. a. Pandey, S. a. Shukla and KK, "Comparison analysis of link prediction algorithms in social network," International Journal of Computer Applications, vol. 111, pp. 27-29, 2015. https://doi.org/10.5120/19624-1502

[2] D. J. a. P. D. C. a. N. C. W. a. R. A. Sanderson, "Graph theory and the analysis of fracture networks," Journal of Structural Geology, vol. 125, pp. 155-165, 2019.

[3] Yaghi, R. I., Faris, H., Aljarah, I., Ala'M, A. Z., Heidari, A. A., & Mirjalili, S. (2020). Link prediction using evolutionary neural network models. In Evolutionary Machine Learning Techniques (pp. 85-111). Springer, Singapore. https://doi.org/10.1007/978-981-32-9990-0_6

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

[4] Jeyaraj, G. T., & Sankar, A. (2019). Extreme learning machine and K-means clustering for the improvement of link prediction in social networks using analytic hierarchy process. International Journal of Enterprise Network Management, 10(3-4), 371-388. https://doi.org/10.1504/ijenm.2019.103162

[5] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512. https://doi.org/10.1126/science.286.5439.509

[6] Hopcroft, J., Lou, T., & Tang, J. (2011, October). Who will follow you back? reciprocal relationship prediction. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1137-1146). https://doi.org/10.1145/2063576.2063740

[7] Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N. V., Rao, J., & Cao, H. (2012, December). Link prediction and recommendation across heterogeneous social networks. In 2012 IEEE

12th International conference on data mining (pp. 181-190). IEEE. https://doi.org/10.1109/icdm.2012.140
[8] Gao, F., Musial, K., Cooper, C., & Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. Scientific programming, 2015. https://doi.org/10.1155/2015/172879
[9] Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1), 17-60.
[10] S. Milgram, "The small world problem," Psychology today, vol. 2, pp. 60-67, 1967
[11] Salihoun, M. (2020). State of Art of Data Mining and Learning Analytics Tools in Higher Education. International Journal of Emerging Technologies in Learning (iJET), 15(21), 58-76. https://doi.org/10.3991/ijet.v15i21.16435
[12] Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. Physica A: statistical mechanics and its applications, 390(6), 1150-1170. https://doi.org/10.1016/j.physa.2010.11.027
[13] Zhu, X., Yang, X., Ying, C., & Wang, G. (2018). A new classification algorithm recommendation method based on link prediction. Knowledge-Based Systems, 159, 171-185. https://doi.org/10.1016/j.knosys.2018.07.015
[14] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 7, 1-30.

6 Authors

Aya Taleb is a Jordanian computer scientist and the main contributor to this work, and was a student at the University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology, Amman (Jordan).
Email: aqrabawi.aya@gmail.com

Prof. Rizik Al-Sayyed is a Jordanian Prof. of Networks, Databases, and Data Science at the University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology, Amman (Jordan).
E-mail: r.alsayyed@ju.edu.jo

Prof. Hamed Al-Bdour is a Jordanian Prof. of Computer systems & Networks at the University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology, Amman (Jordan). E-mail: h.bdour@ju.edu.jo

Article submitted 2021-03-22. Resubmitted 2021-04-16. Final acceptance 2021-04-17. Final version published as submitted by the authors.