# Drug Side Effects Prediction with Random Walks with Restart

Suchanuch PIRIYASATIT
Student ID: 2022280055
*z-yt22@mails.tsinghua.edu.cn*

## I. Abstract

Identification of drug side-effects is critical for the development of drugs as well as for the better understanding of chemical molecules and proteins interactions. In this project, a heterogeneous network of drug and side-effects is constructed based on drug chemical structures, drug protein targets, and associated drugs of side-effects. Random walk with restart is then applied to the heterogeneous network, starting from a query drug, to get the candidate side-effect rankings from the steady-state probability. The quality of the ranking is then evaluated by calculating rank cutoff curve and overall AUC. The obtained ranking is better than a random ranking with AUC of 0.6209 when no prior side-effects information of the query drug is used as the seed nodes and 0.6226 when 10% of known side-effects are used as the seed nodes along with the query drug. This shows a potential in drug side-effects prediction using biological network information. Improvement can be made by including more relevant biological information in the network.

## II. Introduction

Predicting possible side-effects of a drug is essential in the drug developmental process as unexpected side effects can harm patients and even lead to death. Drug safety information is usually obtained by clinical trials which are costly and time consuming. This slows down the drug discovery process and reduces the number of approved drugs in contrast to the significant research efforts [1]. This problem inspires researchers to propose various computation methods to identify drug side effects with the available data. As side effects occurred are considered as disturbances in the form of molecular intercommunication such as protein communication or signal pathways, predicting drug side effects then requires biological network information on the drug compounds, protein targets, and the side effects.

## III. Research Background

The existing methods in predicting drug side effects are based on the assumption that similar drugs have similar chemical and biological characteristics, including chemical structures, protein targets, and side effects [2]. These information are usually taken from various types of biological networks, such as protein-protein interaction networks, drug-target interaction network, and gene regulation network. Not limiting to side effects prediction, these network information along with network analysis methods have enable researchers to make sense of complex molecular systems and identify meaningful associations among molecules [3].

Random walks is one of the state of the art methods in network analysis that relies on guilt by association principle, stating that molecules or genes that are closer in the network tend to have similar properties. It can be thought of as a process of propagating signals through the network, starting from nodes of interest, where the final values of each node in the network reflects its proximity to the seed nodes [4].

Random walk with restart is a modification of the normal random walk by at every time step, a fixed restart probability of walking back to the seed nodes is introduced. This modified random walk is normally used in network propagation as it is able to capture the local neighborhood of the seed nodes as well as the network global structure [4]. As when run the normal random walk for some time, the signal will spread out over the network and the local information of the seed nodes will be lost.

Its applications include the discovery of network modules in a PPI network, where the highest ranked node from the current module seed nodes will be added to the module [5]. In predicting disease causal genes, random walks with restart was applied using known causal genes of similar diseases as node seeds [6]. [7] uses canonical correlation analysis to come up with side-effect seed nodes and applied random walk with restart on the side-effect network to predict drug side-effects.

In this project, we construct a heterogeneous network of side-effect and drug network and apply random walk with restart on the network, starting from the query drug, to get steady-state probability as the ranking of the candidate side-effects.

## IV. Methodology

We first define the walk transition probability of two homogeneous networks (i.e., side effect network and drug network), and then define transition probability going from one network to another (i.e., walk probability from a drug node to a side-effect node and vice versa) to make a heterogeneous network of drug and side-effect.

Side-effect network: We define a side effect similarity matrix $S_s$, where similarity of side-effect $i$ and side-effect $j$

is defined as the jaccard similarity between the set of known drugs associated with each side effect,

$$S_s(i,j) = \frac{|\Gamma_d(i) \cap \Gamma_d(j)|}{|\Gamma_d(i) \cup \Gamma_d(j)|},$$

where $\Gamma_d(i)$ and $\Gamma_d(j)$ are the sets of drug associated with side effect $i$ and $j$, respectively. The similarities is then normalized to take account for the different similarity profiles of different side-effects, $\bar{S}_s = D_s^{-\frac{1}{2}} S_s D_s^{-\frac{1}{2}}$, where $D_s$ is the diagonal degree matrix, holding the node degrees at the diagonal entries.

The side effect transition matrix is then defined based on the side effect similarity matrix with a consideration for a jumping probability $\lambda$ to jump to a drug network if a side effect has a known associated drug.

$$M_{ss}(i,j) = p(s_j|s_i) = \frac{(1-\lambda)\bar{S}_s(i,j)}{\sum_k \bar{S}_s(i,k)}$$

where $A_{ds}$ is the adjacency matrix of drugs and side effects.

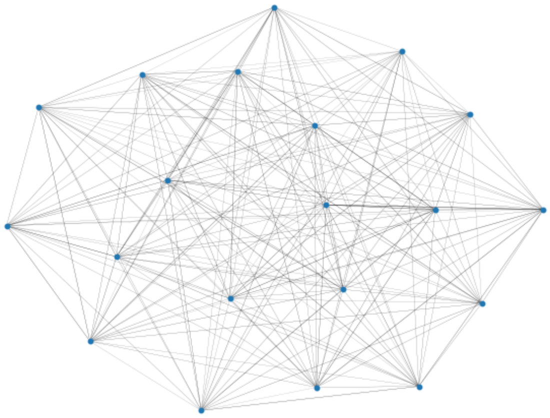

Fig. 1: An example of a side-effect network.

Drug network: The drug similarity matrix $S_d$, following Chen [], is defined to be composed of two types of similarity measures: shared targets similarity, $S_d^t$, and chemical structures similarity, $S_d^c$.

$$S_d^t(i,j) = \frac{|\Gamma_t(i) \cap \Gamma_t(j)|}{|\Gamma_t(i) \cup \Gamma_t(j)|},$$

where $\Gamma_t(i)$ and $\Gamma_t(i)$ are the sets of protein targets associated with drug $i$ and $j$, respectively. Chemical structures similarity, $S_d^c$, is defined to be the dice similarity of Morgan Fingerprint molecule structure of the drug.

$$S_d^c(i,j) = \text{DiceSimilarity}(i,j),$$

Again, each similarity matrix is normalized to account for different similarity profiles of each elements, i.e., $\bar{S}_d^t = D_t^{-\frac{1}{2}} S_d^t D_t^{-\frac{1}{2}}$ and $\bar{S}_c^c = D_c^{-\frac{1}{2}} S_d^c D_c^{-\frac{1}{2}}$, where $D_t, D_c$ are the diagonal matrices. The combined drug similarity matrix is then, $S_{dd} = w_d \bar{S}_d^c + (1-w_d)\bar{S}_d^t$, where $w_d$ is the balancing weight between shared target similarity and chemical structure similarity.

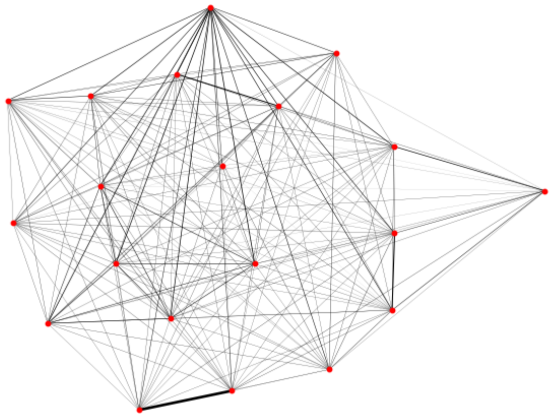

Fig. 2: An example of a drug network. Thicker lines represent higher transition probability between two nodes.

The drug-drug transition matrix is defined as follows

$$M_{dd}(i,j) = p(d_j|d_i) = \begin{cases} \frac{S_{dd}(i,j)}{\sum_k S_{dd}(i,k)}, & \text{if } \sum_k A_{dd}(i,k) = 0 \\ \frac{(1-\lambda)S_{dd}(i,j)}{\sum_k S_{dd}(i,k)}, & \text{otherwise} \end{cases}$$

where $A_{dd}$ is the adjacency matrix of drugs and side effects and again with $\lambda$ probability of jumping to the side-effect network.

Drug - side-effect network: The side effect - drug transition matrix is defined as follows,

$$M_{sd}(i,j) = p(d_j|s_i) = \begin{cases} \frac{\lambda A_{sd}(j,i)}{\sum_k A_{sd}(k,i)}, & \text{if } \sum_k A_{sd}(k,i) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

The drug - side effect transition matrix is defined as follows,

$$M_{ds}(i,j) = p(s_j|d_i) = \begin{cases} \frac{\lambda A_{ds}(i,j)}{\sum_k A_{ds}(i,k)}, & \text{if } \sum_k A_{ds}(i,k) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

The transition matrix, $M$, of the combined heterogeneous network is then defined to be

$$M = \begin{bmatrix} M_{ss} & M_{sd} \\ M_{ds} & M_{dd} \end{bmatrix}$$

Random walk with restart is applied to this network. The iterative probability update is as follows,

$$p_{t+1} = (1-\alpha)M^T p_t + \alpha p_0,$$

where $p_t$ is a vector in which each element $k$ represents the probability of finding the random walker at node $k$ at timestep $t$. At each iteration, the random walker has the restart probability $\alpha$ of walking back to the seed nodes. $p_0$ is the

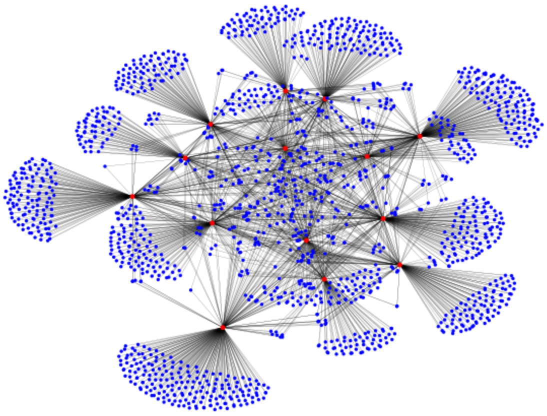

Fig. 3: An example of a drug - side effects network. Red nodes represent drugs and blue nodes represent side-effects.

initial probability of the seed nodes. If a drug is assumed to have no known side-effects, the only seed node is the drug. If a drug is assumed to have some known side-effects, the known side-effects and the drug are the seed nodes with initial probability equally divided among them.

As the network is connected and $M$ is normalized so that the eigenvalues are at most 1 in absolute value, the process is shown to converge to the analytical solution of steady-state distribution $p$ [4],

$$p = \alpha(I - (1-\alpha)M^T)^{-1}p_0$$

This steady-state probability distribution is the final side effects rankings to the query drug in each random walk pass.

## V. Data

Data of side effects of individual drugs and drug-target protein associations are taken from [8]. There are a total of 523 side effects and 241 drugs used in this experiment.

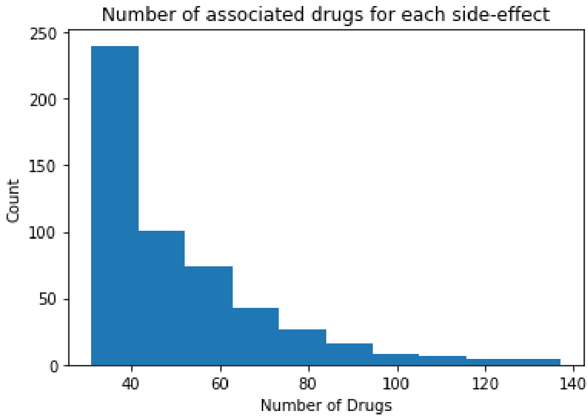

Fig. 4: Histogram of the number of associated drugs for each side-effect.

There are originally over 10,184 side effects which would create a computation time issue when constructing a similarity matrix requiring a computation of all similarity pairs. Thus in this preliminary experiment, side effects with less than 30 associated drugs are removed to save computation time. As a result the side effects are reduced to a total of 523 side effects. The mean of the number of side effects per drug is 50. The histogram of the number of associated drugs for each side-effect is shown in figure 4. The mean of the number of drugs per side-effect is 109. The histogram of the number of associated side-effects for each drug is shown in figure 5.

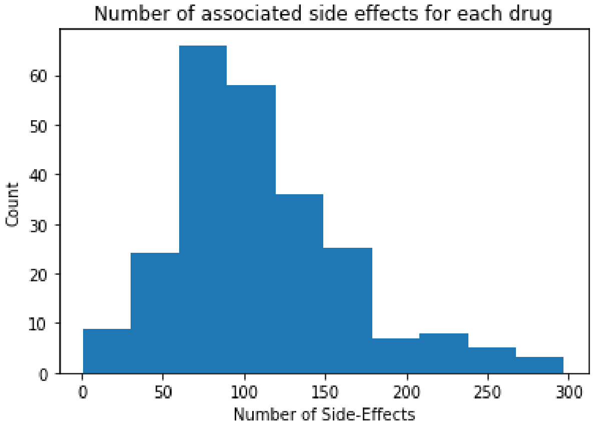

Fig. 5: Histogram of the number of associated side-effects for each drug.

## VI. Experimental Result

We experimented with using the query drug as the only seed node and using the query drug along with some known side effects of the query drug as seed nodes. $\lambda = 0.5, w_d = 0.7, r = 0.2$ are used as initial parameters. Different values of the restart probability is also used to investigate the effects.

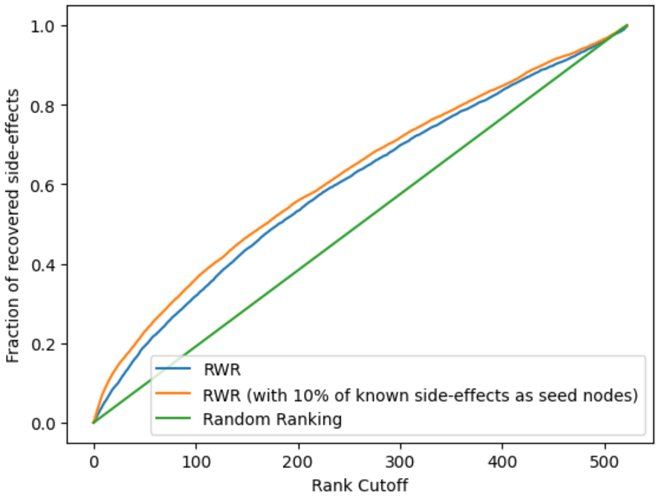

Fig. 6: Rank cutoff curves showing fraction of known side-effects ranked above various cutoffs

Two evaluation methods are used to evaluate the quality of the final side-effect rankings. The first method is calculating the rank cutoff curve. As for each side-effect ranking of a drug, we want the actual associated side-effects to be among the top rankings. The rank cutoff curve plots the fraction of recovered side-effects ranked above various cut-off. More specifically, in each test run of the random walk, known associated test side-effects of a drug are removed from the network edges, hence, there are no known evidence of their associations in the network. The random walk is expected to recover the test side-effects by ranking them in the higher ranks compared to side-effects that are not associated with the drug. After obtaining final ranking from the walk where the test side effects are ranked among all the candidate side effects, the fraction of recovered test side effects ranked above various cutoffs is calculated, considering all the test cases. The result rank cutoff curve is shown in figure 6, where the ranking from using two variations of seed nodes are compared with the random ranking. The rank cutoff of both curves are above the random ranking baseline.

The second method used is by calculating ROC curve and overall AUC of the ranking scores, shown in figure 7. The obtained AUC is at 0.6209 for when no prior side-effects information of the query drug is used as the seed nodes and 0.6226 when 10of known side-effects are used as the seed nodes along with the query drug. The two curves appeared to have no significant difference. By having the AUC of above 0.5, the output ranking from the random walk is better than a random ranking.

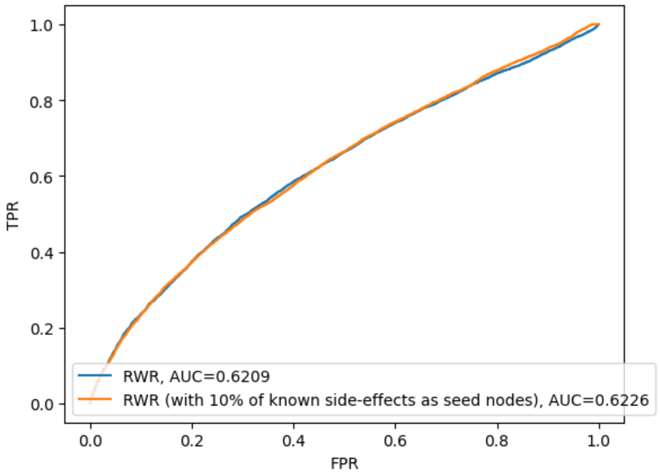

Fig. 7: ROC curve and overall AUC

## VII. Discussion

We experimented with different values of the restart probability, $\alpha$, and similar to [9], found that the result is robust to the restart probability used. Table I shows AUC of different restart probability values. The AUC only slightly increases as the restart probability increases.

The performance evaluated from the rank cutoff curves and overall AUC suggest that the ranking from the random walk is better than a random ranking, showing potential in future side-effects predictions using biological network information. However the performance is still low considering the AUC of 0.6226. This could be a result from the inefficient information on the side-effect nodes used in the network, as the number of shared associated drugs is only used as the similarity measure. Other relevant biological information on the drugs can be further included in the network to improve the random walk performance.

| 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|
| 0.6162 | 0.6209 | 0.6237 | 0.6254 | 0.6264 | 0.6270 | 0.6274 | 0.6275 | 0.6276 |

TABLE I: Comparing different restart probability values and overall AUC.

## VIII. Conclusion

We constructed a heterogeneous network of side-effect and drug network and defined a joint transition probability matrix for the network. We then performed random walk with restart on the network, using query drug and some known side effects as seed nodes to obtain steady-state probability distribution $p$, representing the ranking of side-effect that are most likely to be assiciated with the query drug. The result shows a potential in using biological network to predict drug side-effects. More relevant biological network information can be used to further improve the quality of the random walk ranking.

## References

[1] B. ML., "Druggable targets and targeted drugs: enhancing the development of new therapeutics." *Pharmacology*, vol. 82, 2008.

[2] M.-h. K. Sukyung Seo, Taekeon Lee and Y. Yoon, "Prediction of side effects using comprehensive similarity measures," *BioMed Research International*, vol. 18, 2020.

[3] K. Sachdev and M. Gupta, "A comprehensive review of computational techniques for the prediction of drug side effects," *Drug Development Research*, vol. 81, 04 2020.

[4] I. T. R.-B. e. a. Cowen, L., "Network propagation: a universal amplifier of genetic associations," *Nat Rev Genet*, vol. 18, p. 551–562, 2017.

[5] C. T. . S. A. Macropol, K., "Rrw: repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinformatics*, vol. 10, p. 283, 2009.

[6] P. J. Li Y, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, 05 2010.

[7] S. R. Atias N, "An algorithmic framework for predicting side effects of drugs," *J Comput Biol*, vol. 18, p. 207, 03 2011.

[8] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, p. 457–466, 2018.

[9] X. Chen, M. Liu, and G. Yan, "Drug-target interaction prediction by random walk on the heterogeneous network," *Molecular bioSystems*, vol. 8, pp. 1970–8, 04 2012.