Precision as a measure of predictability of missing links in real networks

Guillermo García-Pérez[1], Roya Aliakbarisani[2,3], Abdorasoul Ghasemi[2], and M. Ángeles Serrano[3]

[1]QTF Centre of Excellence, Turku Centre for Quantum Physics, Department of Physics and Astronomy, University of Turku, FI-20014 Turun Yliopisto, Finland
[2]Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran 1631714191, Iran
[3]Departament de Física de la Matèria Condensada, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain

Abstract

We prove analytically that even the best possible link prediction method for a network ensemble $\mathcal{E}$—generated by assigning undirected links between pairs of nodes i and j with independent pairwise probabilities $\{p_{ij}\}$—yields a limited precision. This suggests an absolute limitation to the predictability of missing links in real networks, due to the irreducible uncertainty arising from the random nature of link formation processes. We show that a predictability limit can be estimated in real networks, and we propose a method to approximate such bound from real-world networks with missing links.

Optimal prediction strategy for a graph ensemble

The optimal strategy of link prediction (OS) for an ensemble $\mathcal{E}$ is the one that maximizes the expected precision

$$\langle Q\rangle = \sum_{G}\sum_{G_{\rm obs}} P(G, G_{\rm obs}) Q(G, G_{\rm obs}, G_{\rm inf}) =$$
$$\sum_{G_{\rm obs}} P(G_{obs}) \sum_{G|G_{\rm obs}\in\mathcal{S}(G)} P(G|G_{\rm obs}) Q(G, G_{\rm obs}, G_{\rm inf})$$

where $G \in \mathcal{E}$ is the original network, $G_{\rm obs}$ is the observed graph, and G_{inf} is the inferred one. We have analytically proved that the optimal strategy is the one generating $G_{\rm inf}$ from $G_{\rm obs}$ by **adding the links according to the connection probabilities $\{p_{ij}\}$ ranked in decreasing order**.

Expected precision of the optimal strategy in network ensembles

1. Rank all possible pair of nodes in $\mathcal{E}$ using $\{p_{ij}\}$ such that $p_l \geq p_{l+1}$
2. Initialize the expected number of correct predictions, $T = 0$; the expected number of non-observed links, $H = 0$; and link index, $l = 1$
3. Repeat

$$T_{\rm new} = T_{\rm old} + P\left(a_l = 1, a_l^{\rm obs} = 0\right) = T_{\rm old} + qp_l$$
$$H_{\rm new} = H_{\rm old} + P\left(a_l^{\rm obs} = 0\right) = H_{\rm old} + 1 - p_l + qp_l$$
$$l = l + 1$$

Unitl $H \approx L = \sum_l qp_l$

4. $\langle Q\rangle = T/H$

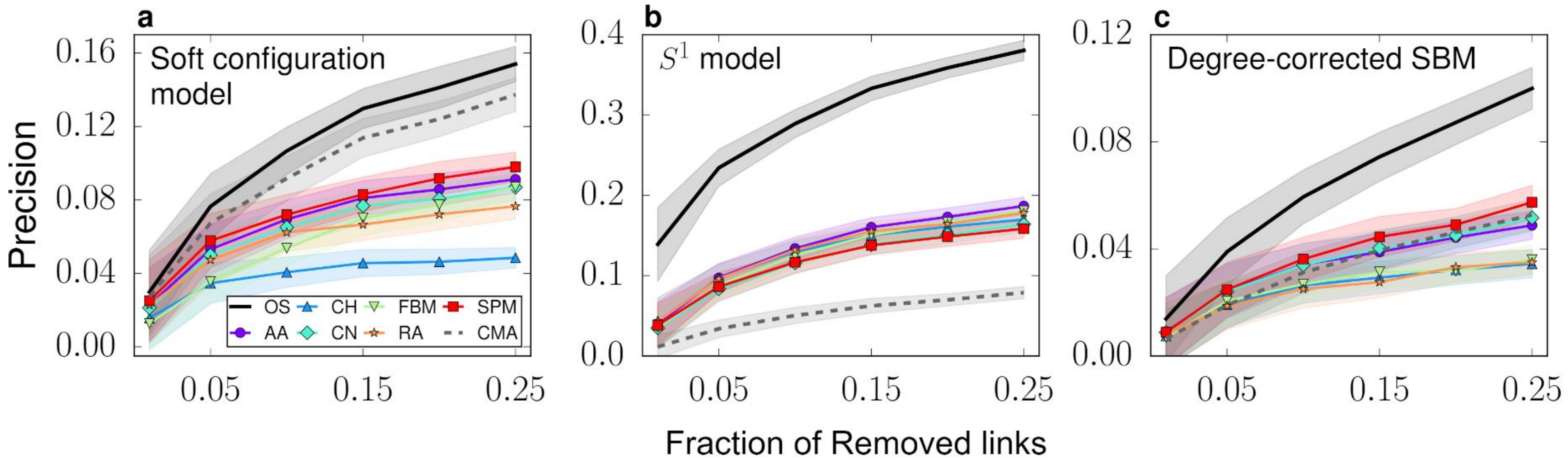

Figure 1: Precision as a function of the fraction of missing links for different link prediction methods on different network ensembles

Inferring the OS predictability curve in fully observed real networks

- Infer the connection probabilities $\{p_{ij}\}$ for a complete real network using a suitable ensemble model—here we use $\mathbb{S}^1$ or dc-SBM.
- The network-specific choice between $\mathbb{S}^1$ and dc-SBM is based on the computed likelihood for network to be generated by each model

$$\mathcal{L} = \prod_{i<j} p_{ij}^{a_{ij}} (1 - p_{ij})^{1-a_{ij}}.$$

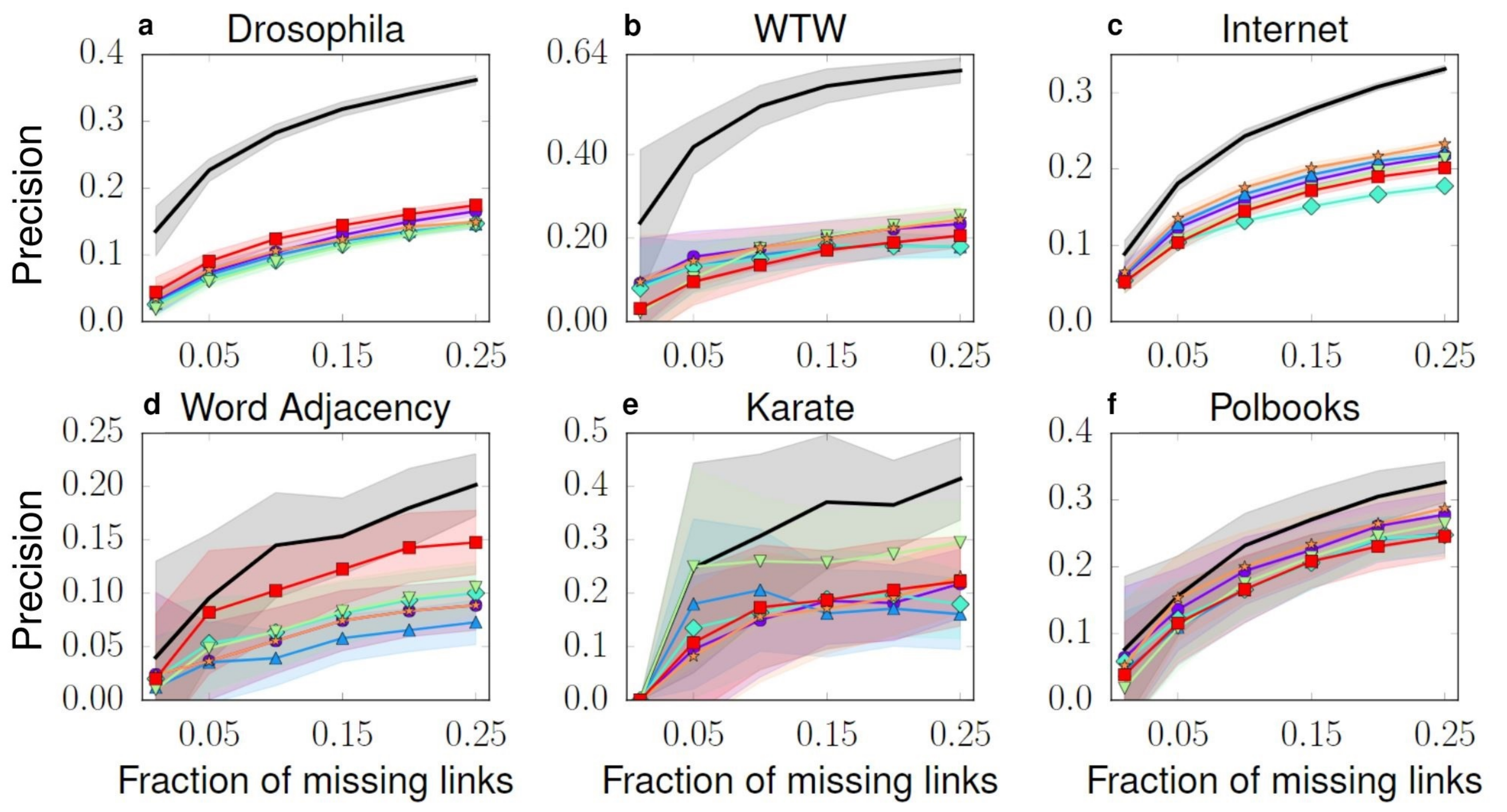

Figure 2: Precision as a function of the fraction of missing links for different link prediction methods on complete real-world networks. Panels **a-c** use $\mathbb{S}^1$ and **d-f** employ dc-SBM model.

Inferring the OS predictability curve in real networks with missing links

1. Select the most suitable network model for $G_{obs}(V, E_{obs})$
2. Fit the model to infer $\{p_{ij}^{\rm obs}\}$, relabel and sort them as $p_{ij}^{\rm obs} \leftrightarrow p_l$ and $p_l > p_{l+1}$
3. Initialize the expected number of non-links in new graphs $\tilde{G} \in \mathcal{S}(G)$, $H = 0$; the expected number of non-links in $\tilde{G}$ that would exist in G, $T = 0$; and link index, $l = 1$
4. Repeat
   (a) if two node pair in l are connected in G_{obs}

$$T_{\rm new} = T_{\rm old} + q$$
$$H_{\rm new} = H_{\rm old} + q$$
$$l = l + 1$$

   (b) if two node pair in l are not connected in G_{obs}

$$T_{\rm new} = T_{\rm old} + q\frac{q_0 p_l}{1 - p_l + q_0 p_l}$$
$$H_{\rm new} = H_{\rm old} + \frac{1 + (qq_0 - 1)p_l}{1 - p_l + q_0 p_l}$$
$$l = l + 1$$

Unitl $H \approx \tilde{L} = qE_{\rm obs}/(1 - q_0)$

5. $\langle Q\rangle = T/H$

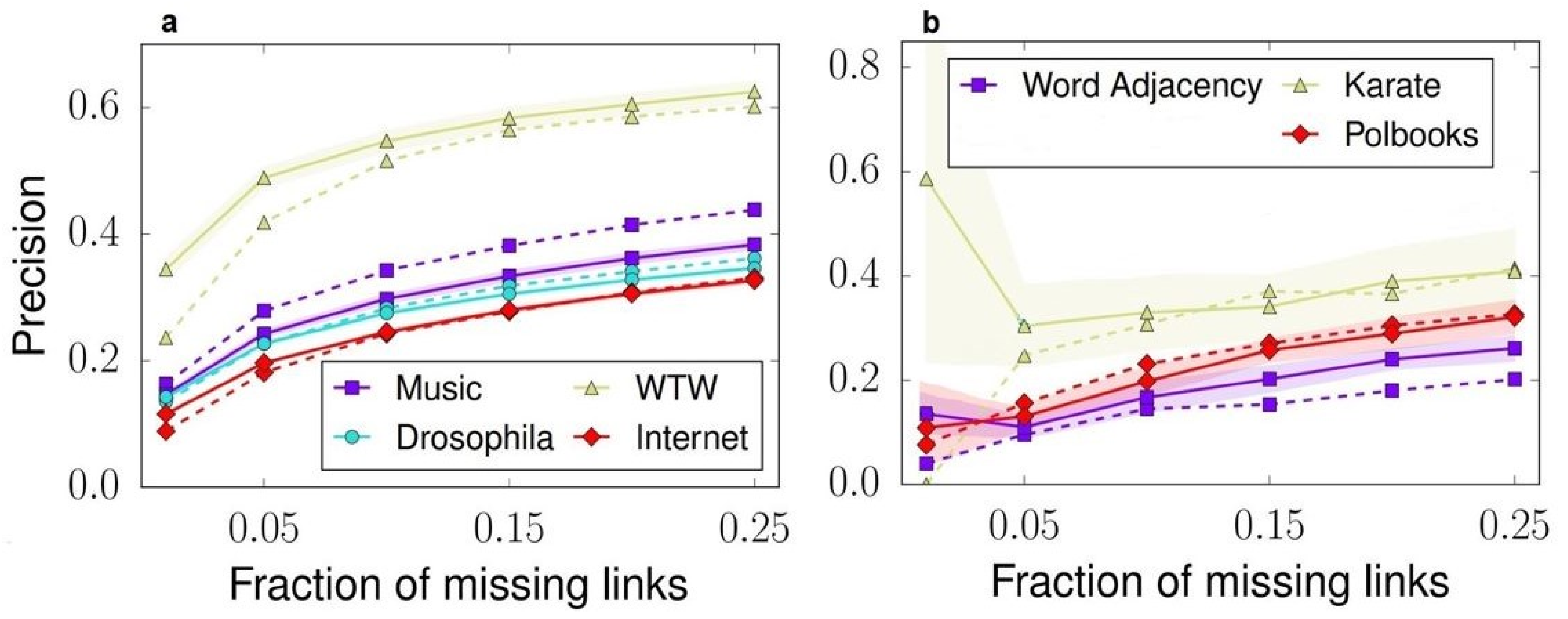

Figure 3: Inference of predictability on real-world networks with missing links

Conclusion

The optimal strategy for link prediction which ranks the likelihood of missing links according to the ensemble connection probabilities outperforms all the link prediction methods on different network ensembles. The OS predictability curve is approximated by fitting proper network model to fully observed real networks and can be used as a benchmark to assess the goodness of link prediction methods. This upper-bound can also be achieved with good accuracy for real networks with missing links using cumulative sequential computation.

Acknowledgment: Guillermo García-Pérez acknowledges financial support from the Academy of Finland via the Centre of Excellence program (Project No.312058 as well as Project No.287750) and from the emmy.network foundation under the aegis of the Fondation de Luxembourg. Roya Aliakbarisani acknowledges the support of the International Affairs department of K. N.Toosi University of Technology. M. Ángeles Serrano acknowledges support from a James S. McDonnell Foundation Scholar Award in Complex Systems, from Ministerio de Ciencia, Innovación y Universidades of Spain Project No. FIS2016-76830-C2-2-P (AEI/FEDER, UE), and from the project “Mapping Big Data Systems: embedding large complex networks in low-dimensional hidden metric spaces” of Ayudas Fundación BBVA a Equipos de Investigación Científica 2017.

G. García-Pérez, R. Aliakbarisani, A. Ghasemi, and M. Á. Serrano, ”Precision as a measure of predictability of missing links in real networks,” *Physical Review E*, vol. 101, no. 5, p. 052318, 2020.