Resumo

O sucesso da teoria dos grafos para descrever sistemas complexos, bem como a onipresença destes, deu muito destaque a elaboração de métodos eficientes para sua analise. No entanto, varias questões continuam em aberto. Uma delas, a qual nos dedicamos neste trabalho, é a obtenção das comunidades presentes nessas redes. Muito embora não exista um consenso formal sobre sua definição, a presença de comunidades vem da ideia intuitiva de que nós formam subgrupos dentro da rede. Neste sentido, muitos algoritmos diferentes foram propostos para identificar tais grupos.

A primeira parte deste trabalho é dedicada ao problema da detecção dessas comunidades. Apresentamos um novo algoritmo, o Surpriser, baseado na maximização da função Surprise, juntamente com um novo benchmark para comparar o desempenho de diferentes algoritmos. Testamos o Surpriser em comparação com sete outros métodos da literatura em três benchmarks diferentes: LFR, RC e o nosso próprio benchmark. Nossos resultados indicam que, entre os métodos com boa resolução, o Surpriser possui o melhor desempenho. Por outro lado, algoritmos com problemas de resolução, como os baseados na Modularidade, tendem a retornar comunidades com tamanhos semelhantes. Em relação aos benchmarks, nossos resultados mostram que o benchmark LFR produz redes com uma forte estrutura de subcomunidades.

Na segunda parte deste trabalho, voltamos nosso foco para as aplicações práticas do algoritmo Surpriser. Criamos e examinamos as redes metabólicas de 8498 organismos diferentes divididos nos três domínios da vida: bactérias, arqueias e eucariotos. Nossas análises revelaram que a estrutura de comunidades desses organismos é composta por muitas comunidades pequenas e de tamanho semelhante. Além disso, demonstramos que a estrutura de comunidades das redes metabólicas surge de pressões evolutivas distintas das que influenciam a distribuição de graus da rede.

Finalmente, utilizamos os atributos dessas redes metabólicas como características para um modelo de classificação taxonômica. A rede neural que treinamos obteve uma precisão de 98% ao classificar os organismos em bactérias, arqueias e eucariotos, no nível de domínio.

Palavras-chaves: Grafos, Redes, Detecção de Comunidades, Benchmarks, Surprise.

Abstract

The success of network science to describe many complex systems and their ubiquitous presence has brought the development of new, more efficient, methods of analysis to the spotlight. However, some problems still remain open, one of which—the focus of our work—is the determination of a network's community structure. Even though there's no consensual formal definition, communities arise from the intuitive idea that nodes form subgroups in larger networks. In this regard, various algorithms have been proposed to identify such groups.

The first part of our work is dedicated to the community detection problem. We introduce a novel algorithm, Surpriser, based on the maximization of the Surprise function, along with a new benchmark for comparing competing algorithms. We tested our Surpriser algorithm against seven other methods from the literature using three different benchmarks: LFR, RC, and our benchmark. Our results demonstrate that among the methods with good resolution, Surpriser performs the best. Conversely, algorithms with poor resolution, such as Modularity-based methods, tend to return communities with similar sizes. Regarding the different benchmarks, we found that LFR produces networks with a strong subcommunity structure.

In the later part of our work, our focus shifted to practical applications of the Surpriser algorithm. We generated and examined metabolic networks for 8498 different organisms spanning the three domains of life: bacteria, archaea, and eukarya. Our analysis revealed that the community structure of these metabolic networks consists of numerous small communities that are mostly evenly-sized. Moreover, we show that the metabolic network's community structure arises from evolutionary pressures distinct from those influencing the network's degree distribution.

Finally, we used the attributes from the metabolic networks as features for a model of taxonomic classification. The neural network we trained achieved an accuracy of 98% in classifying organisms into bacteria, archaea, and eukaryotes at the domain level.

Key-words: Graphs, Networks, Community Detection, Benchmarks, Surprise