User Recognition of Devices on the Internet based on Heterogeneous Graph Transformer with Partial Labels

Yimo Ren[1], Hong Li[1], Peipei Liu[1], Jie Liu[1], Hongsong Zhu[1], and Limin Sun[1]

[1]School of Cyber Security University of Chinese Academy of Sciences Beijing 100049 China

October 24, 2022

Abstract

Recognizing the users of devices, who use IP addresses as unique identity on the Internet, can easily enable numerous security applications. Due to the lot's kinds of device data and large number of missing values, it is difficult to recognize the users of devices well. The community detection methods based on Graph Neural Network (GNN) can integrate multi-source data well to cluster devices into the community with the same user, so as to realize user recognition of devices with higher performance. While existing GNN methods face several issues that the methods on homogeneous graphs could not utilize the multi-source data of devices and most of the methods on heterogeneous graphs need specific knowledge to design meta paths. Also, the web-scale Internet data of devices make it hard for existing methods to learn the representation fully. Finally, a majority of the methods above do not consider the known partial labels in the early stage of training models, not making full use of label information, leading to the effects slightly insufficient. To improve the performance of user recognition, this paper proposes HGT-PL, namely Heterogeneous Graph Transformer with Partial Labels, to calculate the representation of devices on the Internet, on which cluster methods are used to realize user recognition. By using graph transformer, HGT-PL deeply learns node features and graph structure on the heterogeneous graph of devices. By Label Encoder, HGT-PL fully utilizes the users of partial devices from preliminary rules with high confidence. By using cluster methods, the communities with different users are carefully divided and modified. The paper conducts experiments on the web-scale data collected from the Internet and the results show that HGT-PL is able to recognize users of devices more accurately and effectively, with 0.5121 NMI and 0.3554 ARI, compared with existing GNN methods.

Hosted file

`User Recognition of Devices_v1.docx` available at https://authorea.com/users/516762/articles/591483-user-recognition-of-devices-on-the-internet-based-on-heterogeneous-graph-transformer-with-partial-labels