Plan Overview

A Data Management Plan created using DMPTool

Title: Computer Network Traffic

Creator: Krauss Wang

Affiliation: University Of Chinese Academy Of Sciences

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

随着互 网和信息技 的 速 展， 算机网 在全球范 内广泛 用，成 各行各 信息交流和 源共享的基 。然而，伴随而来的网 安全 日益 峻。网 攻 手段不断升 ， 的安全防 措施 以 日益复 的网 威 。特 是 工作站IP的攻 ，因其普遍存在且 常被作 网 攻 的跳板，造成了大量的 意流量。 意流量不 威 着信息系 的安全， 可能 致数据泄露、服 中断等 重后果。因此，及 、准确地 并防御网 流量中的 意行 是当前网 安全 域的重要 。

本 目研究的数据集 Kaggle上下 的“Computer Network Traffic”，

Start date: 07-01-2006

End date: 09-30-2006

Last modified: 05-30-2024

creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Computer Network Traffic

Data Collection

What data will you collect or create?

Kaggle数据网站上的“Computer Network Traffic”，包含 500,000行的网 流量 ，涉及21,000条具体数据。数据集涵盖了10个本地工作站IP在2006年7月1日至2006年9月30日期的网 流量信息，其中一半IP在此期 被攻陷并成 各 僵尸网 的成 。原始数据文件 cs448b_ipasn.csv， processed_cs448b_ipasn_new.csv，其中CSV格式便于 取和共享， 期 和兼容性好。 数据集没有直接可重复使用的 有数据。

How will the data be collected or created?

我 将使用 准化的方法来管理Kaggle上的Computer Network Traffic数据集，包括使用CSV格式、Git 'raw_data/'和'processed_data/'），并通 数据校 、重复采 、 准化数据捕 、数据 入 和同行 等 量保 流程，确保数据的一致性和完整性。

Documentation and Metadata

What documentation and metadata will accompany the data?

数据文件 cs448b_ipasn.csv。

数据集包含4个字段，分 日期、本地IP ASN和流量 次（f）。每个字段的 描述 ：

data: 日期，表示网 流量 的 践，格式 ‘年-月-日’（2006年7月1日~2006年9月30日）；

l_ipn: 本地IP IP 0 9

r_asn: ASN（Autonomous System Number），表示 程ISP的 ，数据 型 整数；

f: 流量 次，表示某日期下从本地IP ASN的流量次数。

Ethics and Legal Compliance

How will you manage any ethical issues?

此公共数据集来源于 http://statweb.stanford.edu/~sabatti/data.html。

Kaggle上的Computer Network Traffic数据集 遵循道德 范，我 已 得了数据的使用 可，并采取了数据匿名化措施，保 参与者的身份。敏感数据将使用加密技 行安全存 和 。我 将遵循机 理委 会的指 ，确保所有数据共享和使用符合 理 准，并得必要的同意。数据保留 将根据 目需求和法律 定 行管理，确保数据的安全和参与者的私保 。

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

数据集遵守CC0：Public Domain

CC0，

Storage and Backup

How will the data be stored and backed up during the research?

我 已 估存 需求，确 当前空 足 ，并 留 外 算用于云服 （如AWS S3）。数据每周 份一次，存 在云端和本地外部硬 ，共三份 份（ 份云端，一份本地）。数据管理 份和恢复，事故 生 确保24小 内恢复。 份数据采用加密技 ， 先使用大学IT

How will you manage access and security?

RBAC）和多因素（MFA）控制 。 作者将通 加密通信渠道（如VPN、SSL）安全 数据。在 收集数据 ，使用加密存 数据，并立即上 至主要系 。此数据集不涉及机密数据，数据来源于公共网站。

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

原始数据集保存在Kaggle中Computer Network Traffic网站中。 了 足合同、法律或 管要求，我 必 保留所有相关的数据，并在 目 束后安全 不再需要的数据。其他数据的保留将基于其潜在的重用价 ，例如 研究 果、 行新研究或用于教学。 保留 至少 5年，以支持未来的研究和教学。我 将 可行且有助于数据共享和保存的格式和内容。准 工作包括整理数据、更新文件格式和生成必要的文档，以确保数据的 期可用性和易于 。

What is the long-term preservation plan for the dataset?

AWS S3）作 份。 目小 成 将 数据的整理和 ，确保数据在 款期限之后也能得到有效管理。 了支持数据的 期保存，我 会 留 和 源 行数据清理、格式更新和文档生成。 些措施确保数据的高可用性和重用价 。

Data Sharing

How will you share the data?

在 目 束后，我 划将 目数据 布到Kaggle网站上，以便更广泛地分享和利用。通 Kaggle，我 可以 数据 外可 ，并吸引数据科学家、研究人 和 界 家参与数据分析和

掘。我 将确保数据在Kaggle上以清晰明了的方式展示，并提供 的数据描述和文档，以帮助用 理解数据的背景和特性。我 将与Kaggle社区共享数据，并鼓励用 参与数据 、 目和 ，以推 数据的 一步研究和 新 用。 做不 可以促 知 共享和 作， 可以 目 来更多的 解和价 ，推 数据科学 域的 展。

Are any restrictions on data sharing required?

CC0：Public Domain

CC0，

Responsibilities and Resources

Who will be responsible for data management?

在我 的 目小 中，数据管理 划的 行由 数据管理的 成 ，他 将确保 数据管理 划的定期 和修 。在我 的 中，每个数据管理活 都有指定的 任人：数据捕 由 目研究人 ，元数据制作由数据管理 家 ，数据 量由 的数据 量 控，存 和 份由IT

What resources will you require to deliver your plan?

在我 的数据管理 划中，我 很幸 地 我 有的硬件 已 足了我 的需求。我 的服 器和存 具有足 的容量和性能，可以支持我 的数据 理和存 需求。因此，我 无需 外投 于硬件 ，并可以将 些 源用于其他关 方面，如数据 量控制或人 培 。 一 我 省了成本，并确保了数据管理 划的 利 施。
