BML : l'outil pour l'analyse de données BGP

Kevin Hoarau, Pierre Ugo Tournoux, Tahiry Razafindralambo

BML : l'outil pour l'analyse de données BGP

Kevin Hoarau[1] et Pierre Ugo Tournoux[1] et Tahiry Razafindralambo[1]

[1] *Université de La Réunion, LIM, France*

Le protocole BGP (Border Gateway Protocol) occupe une place centrale dans l'échange d'information de routage à l'échelle mondiale. Dans la littérature, plusieurs travaux s'intéressent à la détection des anomalies sur le protocole BGP par l'extraction d'attributs statistiques ou de la topologie du réseau BGP à partir de données publiquement disponibles. Cependant, il n'existe pas d'outil générique pour la collecte, l'agrégation, le stockage et la transformation des données BGP. Cet ensemble de tâches fastidieuses freine donc l'émergence de nouveaux travaux et rend difficile la reproduction des résultats. Afin de répondre à ces problèmes, nous avons élaboré et rendu publiquement accessible un outil nommé BML. Cet outil permet de simplifier et d'accélérer la construction de jeux de données BGP. BML se voulant simple d'utilisation et polyvalent, il permet l'extraction des attributs les plus utilisés dans la littérature ainsi que l'ajout de nouveaux types de transformation de données BGP.

Mots-clefs : BGP, Jeu de données, Apprentissage automatique

1 Introduction

Le protocole BGP (Border Gateway Protocol) est le protocole de routage qui permet l'interconnexion de l'ensemble des réseaux qui composent Internet. Les défaillances du protocole, à savoir les anomalies BGP, se produisent pour plusieurs raisons allant des pannes matérielles aux attaques intentionnelles [AMBA16].

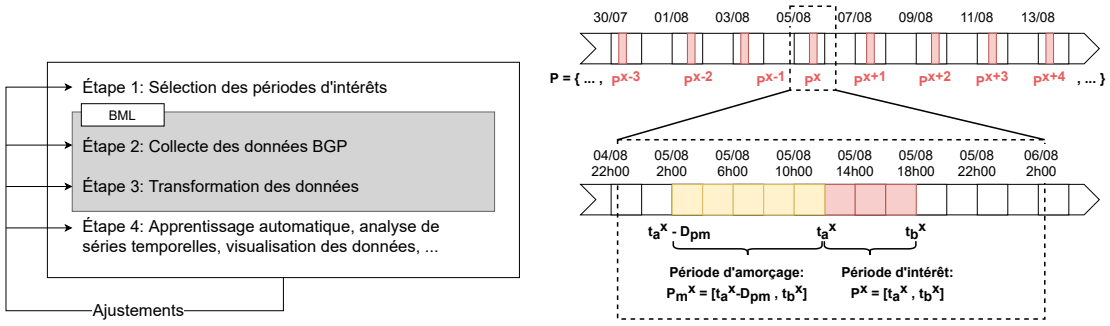
Pour analyser les anomalies BGP, des données collectées et archivées à partir de différents points de collecte sur Internet doivent être traitées pour être exploitables. Cependant, il n'existe aucun moyen systématique et standard de traiter les données BGP et d'en extraire des informations utiles ou permettant leur usage avec des outils d'apprentissage automatique.

De nombreuses approches basées sur des attributs statistiques ou des attributs du graphe BGP ont été développées dans le domaine de la détection d'anomalies BGP. L'étude de différents attributs peut conduire à différentes méthodes de traitement des données qui sont chronophages à mettre en place. Une méthode de traitement standard devrait permettre l'extraction de toutes les caractéristiques importantes et faciliter ainsi l'étude des anomalies BGP.

Les auteurs de [FMBP19] ont développé un outil pour construire des jeux de données BGP. Cependant, cet outil ne peut pas extraire la structure du graphe BGP et les plongements de graphes de BGP utilisés dans [GKHL18, SFPB19]. Leur outil manque également de polyvalence car il nécessite de relancer le long processus de collecte de données afin de modifier les attributs générés.

L'outil présenté dans cet article, appelé BML [HTR21], permet l'extraction d'attributs statistiques de BGP (comme dans [FMBP19]), la structure du graphe BGP et offre aux utilisateurs la possibilité de créer leurs propres attributs. Cet outil automatise les processus de collecte et de transformation des données BGP. BML n'a besoin que d'un ensemble de fenêtres temporelles pouvant être liées à des événements BGP pertinents et d'un ensemble de paramètres qui affectent le compromis entre l'exhaustivité, le temps de collecte et l'empreinte en stockage du jeu de données résultant. Une fois le jeu de données collecté, l'utilisateur peut définir et appliquer plusieurs fonctions de transformation pour exploiter les données selon ses besoins.

Le reste de cet article est structuré comme suit : la section 2 décrit notre outil. La section 3 est dédiée à un exemple de cas d'utilisation de BML. Enfin, la section 4 conclut et discute notre contribution.



(A) Processus de construction d'un jeu de donnée BGP

(B) Collecte de données BGP avec BML

FIGURE 1

2 Notre outil

BML est un outil pour la construction de jeux de données BGP qui a été pensé en priorité pour l'application des techniques d'apprentissage automatique aux données BGP mais son usage ne se limite pas à cette problématique. En effet, BML peut être utilisé à des fins telles que l'analyse statistique des données BGP, ou la visualisation de ces données. La figure 1.A décrit le processus de construction d'un jeu de données BGP dans lequel BML s'inscrit. Ce processus est composé de 4 étapes :

1) Sélection des périodes d'intérêts : Dans BML, une période d'intérêts correspond à une fenêtre temporelle qui intéresse l'utilisateur et qui sera donc inclue dans le jeu de données.

2) Collecte des données BGP : Pour chaque période d'intérêts, BML va automatiquement récupérer les données BGP collectées par différents routeurs BGP dans la fenêtre temporelle correspondante.

3) Transformation des données : Pour cette étape, l'utilisateur doit spécifier une fonction de transformation $T(.)$ qui prend en entrée des données BGP et Δ une période d'échantillonnage. Ainsi, BML va découper les données BGP avec un intervalle de temps Δ et appliquer à chaque segment de données BGP la transformation $T(.)$. Le résultat de cette opération est alors une séquence $[T_1, T_2, \ldots, T_{n-1}, T_n]$.

4) Application : Cette étape consiste en l'utilisation du jeu de données sur la tâche souhaitée. Elle n'est généralement pas définitive car en fonction des résultats obtenus, l'utilisateur voudra généralement effectuer certains ajustements. Un intérêt particulier a donc été porté à la nature itérative de ce processus afin de rendre chaque nouvelle itération à la fois la plus simple et la plus rapide possible.

L'objectif de BML est d'offrir une solution générique pour la construction de jeu de données BGP. De plus, face au volume important des données BGP, une attention particulière a été portée à l'efficacité de la solution afin de rendre possible la collection de jeux de données de grande dimension. Ainsi, lors de sa conception différentes caractéristiques nécessaires ont été identifiées : i) *flexibilité :* ne pas limiter l'utilisateur à un ensemble restreint de transformations possibles, ii) *paramétrable :* BML offre divers paramètres pour la collection et la transformation des données utilisées pour établir un compromis entre la quantité de données à analyser et la charge de calcul, iii) *frugalité :* le téléchargement et l'analyse d'un important volume de données sont des tâche chronophage. L'outil a donc été construit de façon à éviter d'avoir à exécuter cette étape plusieurs fois. iv) *stockage :* pour limiter les besoins en espace de stockage, différents mécanismes ont été mis en place afin de stocker uniquement les données nécessaires à l'exécution des étapes 3 et 4, mais aussi en compressant les données. v) *parallélisation et distribution :* Afin d'exploiter le plus possible la puissance de calcul disponible, BML permet de paralléliser et de distribuer diverses opérations à la fois pour la collecte et la transformation des données.

2.1 Collecte des données BGP

Pour chaque période d'intérêts P^x, l'objectif de BML est de récupérer des données BGP collectées durant une fenêtre temporelle $[t_a^x, t_b^x]$. Pour cela, l'outil s'appuie sur les projets d'archivage de données BGP [HTR21]. Cependant, BPG étant un protocole incrémental, seules des mises à jour de routes sont échangées par les routeurs. Cela, amène donc à un problème de démarrage à froid dans la collecte des données car peu

de données sont disponibles à $t_a^x + \epsilon$, le début de la période d'intérêts. Pour surmonter ce problème, BML permet de définir une période d'amorçage afin de collecter des données BGP avant la période d'intérêts (voir la figure 1.B).

Durant la période d'amorçage, deux modes de fonctionnement sont possibles : la collecte des mises à jour de route uniquement ou la collecte des tables de routage les plus récentes qui sont ensuite augmentées avec les mises à jour de route reçues jusqu'à t_a^x. La 2ème option offre une vue plus complète des routes BGP au détriment du temps de collection des données. Après la collecte des données d'amorçage, les mises à jour de routes capturées durant la période d'intérêts sont récupérées.

La collecte des données BGP peut être ajustée à partir d'un ensemble de paramètres qui doivent être fournis à BML. Parmi ces paramètres de collecte, on retrouve notamment : i) *Projects* : définit le ou les projets d'archivages à utiliser pour la récupération des données, ii) *Collectors* : permet de filtrer les collecteurs à utiliser, iii) *IpVersion* : permet de filtrer les routes à utiliser en fonction de la version du protocole IP, iv) *PrimingPeriod* : définit la durée de la période d'amorçage, v) *UseRibsPriming* et *UseRibsData* : permet de choisir si les tables de routage doivent être utilisées pour la période d'amorçage et/ou période d'intérêts.

2.2 Transformation des données

Une fois les données collectées pour toutes les périodes d'intérêts du jeu de données, BML peut être utilisé pour transformer ces données dans un format spécifié par l'utilisateur. Pour cela, l'utilisateur devra choisir une fonction de transformation déjà existante dans BML ou implémenter sa propre fonction de transformation. Cette fonction de transformation $T(.)$ est ensuite utilisée afin de transformer chaque période d'intérêts en une séquence $T^x = \{T_{t_a^x}, T_{t_a^x+\Delta}, \ldots, T_{t_b^x}\}$. Dans BML une fonction de transformation est définie par $T_t = T(S_t, U_{[t-\Delta,t]}, P)$ où S_t est l'instantané des routes [†] calculé au temps t, $U_{[t-\Delta,t]}$ est la séquence des mises à jour de routes collectées entre $[t-\Delta, t]$ et P un ensemble de paramètres de transformation.

Lors de la transformation d'un jeu de données, un ensemble de paramètres P doivent être fournis à BML. Ces paramètres sont composés de paramètres par défaut tandis que d'autres sont spécifiques à certaines fonctions de transformation. Parmi les paramètres par défaut, on retrouve notamment : la valeur de l'intervalle d'échantillonnage Δ, les collecteurs à utiliser et la version IP des routes à traiter.

Une des caractéristiques principales de BML est sa flexibilité par rapport aux fonctions de transformation qu'il est possible d'implémenter. Cependant, afin de faire économiser du temps aux utilisateurs et d'encourager la reproductibilité des résultats, BML propose par défaut plusieurs fonctions de transformations dont notamment : l'extraction d'attributs statistiques [CLL+18], la reconstruction du graphe BGP [GKHL18] et l'extraction d'attributs du graphe BGP [SFPB19].

3 Cas d'application

Dans cette section, nous illustrons comment BML peut être utile pour l'analyse rapide d'évènements liés à BGP. Pour cela, nous proposons d'étudier la fuite de routes provoquée par Google le 25 août 2017 [‡]. À 3 heures 22 UTC, l'AS 15169 appartenant à Google a accidentellement commencé à annoncer plusieurs routes vers des préfixes qu'il ne possède pas et est ainsi devenu un fournisseur de transit. Cette erreur a causé des perturbations importantes sur Internet et principalement au Japon.

Afin d'analyser cet évènement, BML a été configuré pour collecter des données BGP pour une période d'intérêt de 3 heures à 4 heures UTC. Tout d'abord, nous avons extrait toutes les minutes les 32 attributs statistiques implémentés dans BML. L'ensemble du code pour la collecte et la transformation du jeu de données ne nécessitant que 20 lignes de code Python. De plus, 14 attributs ont également été extraits à partir du graphe BGP. Enfin, un avantage majeur de BML étant la possibilité pour un utilisateur de définir facilement sa propre fonction de transformation des données BGP, nous illustrons cette possibilité en supposant que l'utilisateur souhaite visualiser l'évolution du nombre d'annonces dans lesquelles l'AS 15169 appartenant à Google est présente. Cette nouvelle transformation peut être implémentée en seulement 10 lignes de code avec BML.

†. Un instantané des routes est un format de données utilisé par BML pour représenter de manière concise l'ensemble des routes BGP actives à un instant t.

‡. https ://bgpmon.net/bgp-leak-causing-internet-outages-in-japan -and-beyond/

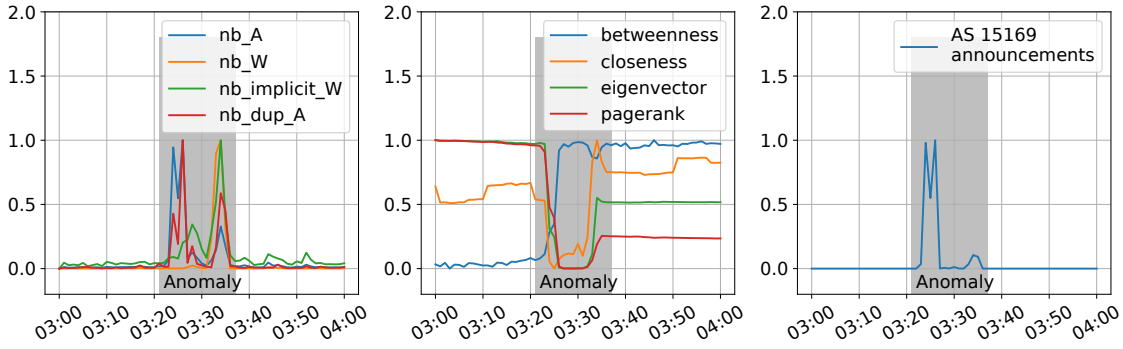


FIGURE 2 : Différents attributs extraits par BML durant une anomalie BGP

Par contrainte d'espace, la figure 2 montre l'évolution de 4 attributs statiques, 4 attributs du graphe BGP et le résultat de la fonction de transformation définie ci-dessus. Ces résultats démontrent l'intérêt de BML pour la surveillance et l'étude à posteriori des anomalies BGP.

4 Conclusion et Perpectives

Cet article présente BML [§], un outil de collecte de données BGP. BML facilite le développement de modèles d'apprentissages automatiques en simplifiant la construction de jeux de données BGP et en permettant aux chercheurs de se concentrer sur l'élaboration et l'optimisation de leurs modèles. Cet outil permet de générer 32 des attributs statistiques les plus utilisés dans la communauté et 14 des attributs du graphe BGP. Les utilisateurs peuvent également implémenter leurs propres transformations à appliquer aux données afin d'extraire des représentations adaptées à leurs modèles. Nous avons illustré la plus-value de BML sur l'analyse de l'anomalie BGP induite par la fuite de routes de Google en 2017. Des outils tels que BML peuvent également aider les chercheurs à partager leurs jeux de données et promouvoir la reproductibilité des résultats obtenus. Il est important de souligner que BML est le premier outil générique pour la collecte, l'agrégation, le stockage et la transformation des données BGP.

Références

[AMBA16] Bahaa Al-Musawi, Philip Branch, and Grenville Armitage. Bgp anomaly detection techniques : A survey. *IEEE Communications Surveys & Tutorials*, 19(1) :377–396, 2016.

[CLL+18] Min Cheng, Qing Li, Jianming Lv, Wenyin Liu, and Jianping Wang. Multi-scale lstm model for bgp anomaly classification. *IEEE Transactions on Services Computing*, 2018.

[FMBP19] Paulo Fonseca, Edjard S Mota, Ricardo Bennesby, and Alexandre Passito. Bgp dataset generation and feature extraction for anomaly detection. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2019.

[GKHL18] Palash Goyal, Nitin Kamra, Xinran He, and Yan Liu. Dyngem : Deep embedding method for dynamic graphs. *arXiv preprint arXiv :1805.11273*, 2018.

[HTR21] Kevin Hoarau, Pierre Ugo Tournoux, and Tahiry Razafindralambo. Bml : an efficient and versatile tool for bgp dataset collection. In *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2021.

[SFPB19] Odnan Ref Sanchez, Simone Ferlin, Cristel Pelsser, and Randy Bush. Comparing machine learning algorithms for bgp anomaly detection using graph features. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, pages 35–41. ACM, 2019.

§. https ://github.com/KevinHoarau/BML