Introduction aux Processus Stochastiques

Projet Chaîne de Markov en temps discret

Analyse de la propagation d'un virus informatique par chaîne de Markov

Profs.: Céline Esser et Pierre Geurts

Année académique 2023-2024

Ce travail est à réaliser par groupe de 2 étudiants. Le rapport et le code source sont à remettre via Gradescope pour **le vendredi 17 mai 2024 à 23h59** au plus tard.

Contexte général et objectifs

Dans ce projet, on se propose de modéliser la propagation d'un virus au sein d'un réseau informatique à l'aide de processus de Markov en temps discret.

Le projet est découpé en quatre étape. On considérera d'abord "sur papier" le cas très simplifié d'un réseau à deux serveurs (Section 1). On étudiera ensuite deux techniques de simulation de la chaîne pour traiter un nombre de serveurs plus important : une simulation au niveau des serveurs (Section 2) et une simulation plus macroscopique (Section 3). Ces deux techniques seront finalement combinées pour simuler un réseau plus conséquent (Section 4). Décrivons d'abord le modèle général de propagation de virus considéré.

Modéle général de propagation du virus. On suppose un ensemble de N serveurs pouvant être dans trois catégories par rapport à un virus informatique [1] :

- Vulnérable (V) : un serveur non encore infectés mais susceptibles de devenir infecté par le virus
- Infecté (I) : un serveur infecté précédemment par le virus et toujours infectieux.
- Protégé (P) : un serveur infectés précédemment mais maintenant débarrassé du virus et non infectieux et protégé (momentanément) d'une nouvelle infection.

Un serveur peut passer de l'état I à l'état P par exemple suite à l'action d'un antivirus.

Ces serveurs sont connectés entre eux en réseau et on souhaite étudier dans ce projet la propagation du virus au niveau des serveurs en prenant en compte ces connexions. Le

1. Ce modèle suit le modèle compartimental de type SIR. Voir par exemple `https://fr.wikipedia.org/wiki/ModÃÍles_compartimentaux_en_ÃľpidÃľmiologie`

réseau sera modélisé par un graphe (non dirigé) reliant ces serveurs, représenté par sa matrice d'adjacence $W \in \{0,1\}^{N\times N}$. Un élément $W_{i,j}$ vaudra 1 si les serveurs i et j sont connectés entre eux (et donc susceptibles de se transmettre directement le virus), 0 sinon [2].

Etant donné ces hypothèses, le modèle (stochastique) de propagation en temps discret proposé est le suivant :

— Un serveur infecté au temps t a une probabilité β d'infecter au temps $t+1$ chacun des serveurs vulnérables auquel il est connecté dans le réseau

— Un serveur infectieux au temps t a une probabilité μ que l'antivirus détecte le virus, et donc de devenir protégé, au temps $t+1$. Le fait qu'un serveur passe dans l'état P au temps $t+1$ ne l'empêche pas de pouvoir infecter un autre serveur au temps t selon la première règle.

— Un serveur protégé au temps t a une probabilité α de redevenir vulnérable au temps $t+1$.

La probabilité β modélise le fait qu'un serveur ne communique pas nécessairement avec tous les serveurs auxquels il est connecté à chaque pas de temps et qu'une connexion ne va pas nécessairement mener à une transmission du virus. μ^{-1} et α^{-1} représentent le nombre moyen de pas de temps nécessaires respectivement à la détection et l'éradication du virus par l'antivirus et à la perte de protection d'un serveur (par exemple suite à un arrêt de l'antivirus par négligence de l'utilisateur ou à son obsolescence par rapport à une nouvelle forme du virus).

Tel que décrit, le processus est un processus de Markov en temps discret. Les états de la chaîne correspondante sont représentés par les catégories, parmi 3, de l'ensemble des N serveurs du réseau. Le nombre d'états de la chaîne est donc 3^N.

Le modèle est inspiré du modèle étudié dans cet article [1], dont nous chercherons plus bas à reproduire certains résultats.

1 Modèle(s) exact(s) à deux serveurs

Dans un premier temps, on se propose d'étudier "sur papier" le modèle exact dans le cas où il n'y a que deux serveurs, connectés entre eux. On supposera qu'initialement un des deux serveurs est infecté et l'autre est vulnérable, les deux serveurs ayant la même probabilité d'être infectés.

Questions. Répondez aux questions suivantes :

1. Justifiez que le modèle proposé est bien un processus de Markov en temps discret. Déterminez les états de la chaîne correspondante et représentez le graphe de transition associé à cette chaîne.
2. Caractérisez de la manière la plus précise possible cette chaîne. Est-elle (a)périodique, irréductible, régulière, absorbante ?
3. En fixant β, μ et α respectivement à 0.5, 0.1, 0.05, calculez et tracez sur un graphe l'évolution en fonction du temps du nombre moyen de serveurs (entre 0 et 2) dans chacune des trois catégories (V, I, et P).

2. On supposera par la suite que $W_{i,i} = 0$.

4. Calculez sur base de la matrice de transition avec les mêmes valeurs de paramètres qu'au point précédent le temps moyen (en pas de temps) nécessaire à la disparition totale du virus (plus aucun serveur infectieux). Discutez (en l'illustrant) l'impact des paramètres β, μ, et α sur ce temps.
5. Considérons maintenant le cas où le premier serveur est en fait connecté avec l'extérieur et a donc une probabilité δ (qu'on supposera invariante dans le temps) d'être infecté à chaque pas de temps (s'il est vulnérable).
   (a) Modifiez le graphe de transition pour prendre en compte δ et caractérisez la nouvelle chaîne de Markov ainsi obtenue.
   (b) En supposant $\delta = 0.05$, quel pourcentage de temps chacune des deux serveurs passeront-ils dans l'état infectieux en régime stationnaire ?
6. Dans le cas où $\delta = 0$, les deux serveurs sont indistinguables. Si on ne s'intéresse qu'à l'évolution du nombre de serveurs dans chacune des catégories, V, I ou P, il est possible de modéliser le système par une chaîne de Markov alternative dont les états sont identifiés par trois variables V_t, I_t et P_t donnant le nombres de serveurs (ici, 0, 1, ou 2) dans chaque catégorie à chaque pas de temps t.
   (a) Déterminez les états de cette nouvelle chaîne de Markov et représentez son graphe de transition.
   (b) Reproduisez les expériences des sous-questions 3 et 4 avec cette chaîne et vérifiez que vous obtenez bien les mêmes résultats qu'avec la chaîne précédente.

2 Simulations au niveau des serveurs et seuil d'épidémie

Calculer explicitement la matrice de transition comme dans la section précédente n'est possible que pour des valeurs de N faibles, le nombre d'états devenant rapidement trop élevé. Il est néanmoins toujours possible d'estimer les mêmes courbes et statistiques que dans la section précédente en se basant sur des simulations de la chaîne de Markov. Pour faire cela, on vous demande d'écrire dans cette section un programme permettant de générer une réalisation aléatoire de la chaîne, représentée par l'évolution au cours du temps de l'état des N serveurs, étant donné une matrice d'adjacence W modélisant le réseau et des valeurs de β, μ et α fixées a priori. Sur base de réalisations, vous devrez être capables de mesurer à chaque pas de temps le nombre moyen de serveurs dans les trois classes (V, I, et P) et de calculer le temps nécessaire à la disparition du virus.

Seuil d'épidémie. Dans l'article [1], dans le cas où $\alpha = 1.0$ (un serveur protégé redevient directement vulnérable [3]), les auteurs montrent que si le rapport $\frac{\beta}{\mu}$ est plus petit que $\frac{1}{\lambda^*_W}$, où λ^*_W est la plus grande valeur propre de la matrice W, alors le nombre de serveurs infectés décroîtra rapidement vers 0. Dans le cas contraire, le nombre de serveurs infectés mettra un temps très long avant de s'annuler (l'épidémie est persistante).

3. Strictement, dans notre modèle, même si $\alpha = 1.0$, un serveur passe au minimum un pas de temps dans l'état P, alors que l'état P n'existe pas dans [1]. Cette différence ne remet cependant pas en question l'existence d'un seuil d'épidémie.

Données. Pour répondre aux questions ci-dessous, une matrice d'adjacence W^{sf} (fichier `Wsf.txt` ou `Wsf.npy`) vous est fournie, représentant un graphe "scale-free" défini sur 1000 serveurs comportant 1996 arêtes.

Questions. Répondez aux questions suivantes dans le rapport :

1. En vous mettant dans les mêmes conditions qu'aux sous-questions 3 et 4 de la section 1 (c'est-à-dire un modèle à deux serveurs connectés entre eux et les mêmes valeurs de paramètres), générez un nombre suffisant de réalisations de la chaîne de Markov et reportez sur un graphe l'évolution du nombre moyen de serveurs dans les trois catégories. Calculez également la moyenne du temps de disparition des serveurs infectés sur ces réalisations. Vérifiez que ces résultats confirment les résultats de la section précédente.
2. Calculez les mêmes courbes pour le graphe W^{sf} fourni en utilisant des valeurs $\beta = 0.3$, $\mu = 0.4$ et $\alpha = 0.0$ et en supposant qu'initialement 0.5% des serveurs (choisis au hasard) sont infectés. Faites des moyennes sur un nombre suffisant de réalisations pour que vos estimations soient stables.
3. En utilisant les autres paramètres fixés comme au point précédent, étudiez l'impact du paramètre α du modèle. Testez des valeurs relativement faibles croissantes de α et observez l'impact sur la convergence du nombre de serveurs infectés. Discutez ces résultats en fonction de ce que prédit la théorie.
4. Réaliser une expérience pour vérifier l'existence du seuil d'épidémie pour la matrice W^{sf} fournie (dans le cas où $\alpha = 1.0$). Par exemple, fixez β à une valeur faible et μ ensuite tel que $\frac{\beta\lambda_W^*}{\mu}$ soit de l'ordre de 0.8 et ensuite de l'ordre de 2 et tracez dans les deux cas l'évolution du nombre moyen de serveurs dans l'état I sur une durée suffisamment grande (≥ 1000), en partant de l'état initial où tous les serveurs sont infectés. Comme dans l'article (figure 5), tracez les graphes en utilisant une échelle logarithmique sur les deux axes. Discutez les courbes obtenues et leur accord avec la théorie.

3 Simulations macroscopiques

L'utilisation de simulations numériques permet de traiter des plus grands graphes mais reste très lourde dans le cas où le nombre de serveurs est très grand. Dans le cas d'un réseau complètement connecté, les serveurs deviennent cependant indistinguables et il est alors possible d'utiliser la même idée qu'à la question 6 de la section 1, c'est-à-dire simuler directement l'évolution du nombre de serveurs dans chaque catégorie, plutôt que de maintenir l'état de tous les serveurs.

Questions. Dans votre rapport, répondez aux questions suivantes :

1. Si on note V_t, I_t, et P_t le nombre de serveurs dans chaque catégorie au temps t, expliquez comme générer (efficacement) de nouvelles valeurs V_{t+1}, I_{t+1} et P_{t+1} selon le modèle de propagation du virus.
2. Implémentez un simulateur de la chaîne basée sur ce principe et comparez les nombres moyens de serveurs par catégorie obtenus avec ce simulateur avec ceux obtenus avec la chaîne de Markov à la question 3 de la section 1 et avec le simulateur de la question 1 de la section 2.

3. Effectuez une simulation avec un graphe de taille $N = 1000$ complètement connecté et comparez le résultat obtenu exactement dans les mêmes conditions avec le simulateur de la section précédente. Choisissez pour cette expérience des valeurs de paramètres représentatives.

4 Simulation à plus grande échelle

Dans cette dernière partie, on propose d'utiliser les outils des sections précédentes pour aborder un scénario un peu plus réaliste. Un nouveau virus est introduit sur un réseau connectant différentes entreprises et on aimerait savoir à quelle vitesse et quels dégâts il est susceptible de faire au niveau du réseau et d'une entreprise en particulier.

Modèle. Soit N_e entreprises disposant chacune d'un parc de serveurs. On supposera que les N_i^s serveurs de chaque entreprise (pour $i = 1, \ldots, N_e$) sont complètement connectés les uns avec les autres par un réseau interne. Les entreprises sont par ailleurs connectées entre elles par un réseau représenté par une matrice d'adjacence W de taille $N_e \times N_e$. On supposera le modèle de propagation suivant inspiré des développements précédents :

— Chaque serveur infecté au temps t a une probabilité β_1 d'infecter au temps $t+1$ chacun des serveurs vulnérables de son entreprise et une probabilité β_2 d'infecter au temps $t+1$ chacun des serveurs vulnérables des entreprises auquel son entreprise est connectée.

— Chaque serveur infecté au temps t a une probabilité μ d'être protégé au temps $t+1$.

On supposera pour cette question qu'un serveur protégé le reste indéfiniment ($\alpha = 0$). L'utilisation de deux paramètres β_1 et β_2 permet de prendre en compte des différences potentielles de taux d'infection à l'intérieur et à l'extérieur d'une entreprise. On supposera par la suite que $\beta_1 > \beta_2$, les connexions en interne étant moins sécurisées et/ou plus fréquentes que vers l'extérieur.

Scénario, données et paramètres. On vous fournit pour répondre aux questions ci-dessous, un graphe W^e défini sur 1000 entreprises et un vecteur N^s contenant le nombre de serveurs de chaque entreprise (entre 1 et 100, pour un total de 49827 serveurs) (fichiers `We.txt` et `Ne.txt` sur Ecampus). On supposera que $\mu = 0.2$, et que $\beta_1 = 2 * \beta_2 = 0.006$. On sait que le virus a vraisemblablement été introduit initialement sur un serveur de la société 0, qui est celle qui est la plus connectée dans le réseau, dans le but de faire le plus de dégâts possible. L'entreprise 220 est en fait le SEGI qui possède 51 serveurs. Les délibérations sont proches et le directeur du SEGI voudrait éviter à tout prix un trop grand nombre de machines infectées simultanément. Le directeur estime que les délibérations pourront se dérouler sans accroche si moins de 20% des serveurs sont infectés simultanément.

Questions. Dans votre rapport, répondez aux questions suivantes :

1. Implémentez le modèle de simulation proposé. Expliquez brièvement son principe général dans le rapport. Vu le nombre total de serveurs très élevé, pour que les simulations puissent se faire en un temps raisonnable, il faudra combiner les simulations macroscopiques de la section 3 avec les simulations dirigées par le graphe de la section 2.

2. Sur base de votre simulateur, tracez les courbes d'évolution du nombre total moyen (sur 25 simulations au moins) de serveurs dans chaque catégorie en fonction du temps pour le réseau entier et pour les entreprises 0 (d'où est partie l'épidémie) et 220 (le SEGI).
3. Afin de rassurer le directeur du SEGI, déterminer la probabilité qu'il y ait au moins 20% de machines infectées simultanément au SEGI (estimée par le pourcentage de simulations pour lesquels le nombre maximum de serveurs simultanément infectés est strictement supérieur à 10).
4. Cette probabilité étant trop élevée, le directeur fait appel aux spécialistes en cybersécurité de l'université (Professeurs Mathy et Donnet) pour améliorer l'antivirus utilisé au SEGI, ce qui permettrait d'augmenter la valeur de μ (pour les serveurs du SEGI uniquement). Déterminez quelle devrait être la valeur de μ du nouvel antivirus pour que la probabilité de dépasser 10 machines infectées soit proche de zéro.

Resources

L'énoncé du projet et les fichiers mentionnés ci-dessus sont disponibles sur Ecampus (rubrique "Projet").

Soumission

Vous devez nous fournir un rapport *au format pdf* contenant vos réponses, concises mais précises, aux questions posées ainsi que le code que vous avez utilisé pour y répondre. Le rapport et le code doivent être soumis de manière séparée sur Gradescope (code cours : EJZBJX). Les deux doivent être soumis pour **le vendredi 17 mai 2024 à 23h59** au plus tard.

Pour vos expériences, vous pouvez utiliser le langage de programmation que vous souhaitez (avec notre accord, si ce n'est pas Mathematica, R, Python, Java ou C). Quel que soit le langage, on vous demande d'implémenter les fonctions de simulations par vous-mêmes. Vous ne pouvez pas utiliser une boîte à outils existante qui ferait exactement ce qui est demandé. En cas de doute, contactez nous.

Références

[1] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4), January 2008.