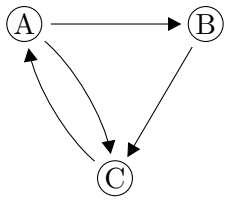
INTRODUCTION

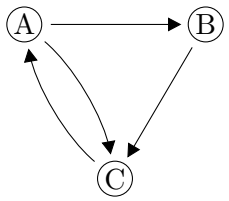
On cherche un algorithme pour déterminer un *ordre d'importance* dans un réseau.

Exemple 5.1. Une ligue de football comprend trois équipes: A, B et C. En tout, ils ont joué quatre jeux: A a perdu contre B et C, B a perdu contre C, et C a perdu contre A. Ceci se résume dans un graphe où $X \to Y$ indique que X et Y ont joué une partie, et que Y a gagné.



Si il y aurait e une équipe qui avait gagné contre toutes les autres, on pourrait facilement l'identifier comme la meilleure. Mais ce n'est pas le cas (et en générale ce ne serait pas le cas). Comment choisir? □

Exemple 5.2. Un internet comprend trois pages web: A, B et C (c'est un exemple simplifié...). Les liens sont indiqués dans le graphe suivant, où $X \to Y$ indique que la page X fait un lien vers la page Y.



Déterminer la page web "le plus important"; mieux, donner un ordre aux pages. □

ÉQUIPES

Au lieu de déterminer la meilleure, identifions plutôt un "score" pour chaque équipe. L'ordre des équipes correspondra à l'ordre des scores. La meilleure équipe sera celle avec la plus grande score, la deuxième équipe sera celle avec la deuxième score, etc. Comment déterminer les scores?

On pose w_A, w_B et w_c les trois scores. Chaque victoire devrait avancé le score, mais de combien? On pourra simplement dire que le score est le nombre de victoires de l'équipe qui donnera $w_A = 2$, $w_B = 1$ et $w_C = 1$. Même sur ce petit exemple, on voit déjà des difficultés: on aura que A est meilleur et que B et C sont égales, pourtant C a gagné contre la meilleure équipe et B n'a pas. Les équipes B et C sont classées comme égales, mais la victoire de C été "plus importante que celle de B. On pose donc plutôt que le score devrait être égale à la

somme des scores de toutes les équipes qui ont été défaites. Pour des raisons techniques, on permet un facteur multiplicatif, donc on pose

$$w_X = \alpha \,(\text{somme do toutes les } W_Y \text{ où } Y \text{ a perdu contre } X) \tag{5.1}$$

Le graphe de la ligue de l'exemple 5.1 peut s'écrire en termes de matrice, òu $A_{ij} = 1$ si $j \to i$, c'est-à-dire si l'équipe j a perdu contre i. On peut aussi mettre les scores en vecteur. Pour l'exemple 5.1 on obtient alors

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_A \\ w_B \\ w_C \end{bmatrix}$$

Alors la condition équation (5.1) s'écrit comme $\mathbf{w} = \alpha A\mathbf{w}$. En autres mots, on a une équation de valeur et vecteur propre $A\mathbf{w} = \frac{1}{\alpha}\mathbf{w}$ avec valeur propre α et vecteur propre $\mathbf{w}$. Les scores sont exactement les composants du vecteur propre.

Comment choisir le vecteur propre? Il y a peut-être plusieurs. On s'inspire des idées des chaînes de Markov.

Exercice 5.3. Est-ce que la matrice A ci-haut est la matrice de transition d'une chaîne de Markov? □

Bien que la réponse à l'exercice précédent est "non", on a encore un espoir n'est pas perdu, car le théorème 3.14 reste encore valide.

Théorème 5.4. *Soit A une matrice avec toutes les composantes non-négatives. Si le graphe correspondante est fortement connexe et apériodique alors la valeur propre dominante λ_1 est positive et de multiplicité un, toute autre valeur propre $|\lambda_j| < \lambda_1$ et tout composant du vecteur propre correspondant à λ_1 est positif.* □

Le théorème 3.14 est le cas spécial de celui-ci où la valeur propre dominante est égale à 1.

Donc pour résoudre la question de l'exemple 5.1, on devrait calculer le vecteur propre dominante, et *c'est exactement l'ordre d'importance des équipes.*

On détermine avec un peu de calcul que la valeur et vecteur propre dominante sont

$$\lambda_1 \approx 1,32 \qquad \mathbf{v}_1 \approx \begin{bmatrix} 1 \\ 0,75 \\ 1,32 \end{bmatrix}$$

Donc l'équipe C est la meilleure, suivie de A, suivie de B.

On se rappelle que calculer toutes les valeurs et vecteurs propres est une tâche onéreuse, surtout pour des grandes matrices. Par contre, on ne cherche qu'une seule chose: le vecteur propre dominante (on pourrait presque dire vecteur d'état stationnaire sauf que...).

On s'inspire donc d'une technique qu'on a vu pour les chaînes de Markov: on considère les *puissances* de A. Si on prend k suffisamment grande, alors les colonnes de A^k seront toutes des multiples de ce vecteur propre dominante (approximativement). On se rappelle que pour une matrice stochastique régulière P, les colonnes de P^k sont approximativement toutes égales, et égales au vecteur d'état stationnaire (pour k suffisamment grand). Ici, la matrice n'est pas stochastique qui donne qúon aura des *multiples* du vecteur propre dominante.

En calculant (par ordinateur bien sûr!):

$$A^{10} = \begin{bmatrix} 7 & 4 & 5 \\ 5 & 3 & 4 \\ 9 & 5 & 7 \end{bmatrix} \qquad A^{20} = \begin{bmatrix} 114 & 65 & 86 \\ 86 & 49 & 65 \\ 151 & 86 & 114 \end{bmatrix}$$

Exercice 5.5. Vérifier si les colonnes de A^{10} sont toutes des multiples du vecteur propre $\begin{bmatrix} 1 \\ 0{,}75 \\ 1{,}32 \end{bmatrix}$ (e.g., en divisant chaque colonne par sa première valeur). Faire de même pour A^{20}. □

PAGES WEB

On voit que l'exemple 5.2 est essentiellement la même que l'exemple 5.1. Une "perte" est maintenant un "lien vers", mais dans un sens c'est la même chose. Dans les deux exemples, les flèches indiquent la progression: soit vers la meilleure équipe, soit vers la prochaine page.

Il y a une différence technique. Dans la ligue, chaque jeux a une certaine influence: l'information totale est le nombre de jeux. Mais sur l'internet, un page qui fait plusieurs liens donne moins d'importance à chacun. La solution c'est de créer une chaîne de Markov. Dans chaque page, on accorde une probabilité égales à chacun des liens. Ceci est équivalent à diviser chaque colonne de A par sa somme.

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \qquad \longrightarrow \qquad P = \begin{bmatrix} 0/2 & 0/1 & 1/1 \\ 1/2 & 0/1 & 0/1 \\ 1/2 & 1/1 & 0/1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix}$$

La matrice P représente la matrice de transition du "surfeur aléatoire": à chaque page web, il choisit par hasard un des liens et le suit. Au long terme, les surfeurs aléatoires seront décrits par le vecteur d'état stationnaire.

On cherche donc le vecteur d'état stationnaire de P (pour une matrice stochastique, le vecteur d'état stationnaire et le vecteur propre dominante sont exactement la même chose). Quelques calculs donnent la valeur et vecteur propre dominante.

$$\lambda_1 = 1 \qquad \mathbf{v}_1 = \begin{bmatrix} 0,4 \\ 0,2 \\ 0,4 \end{bmatrix}$$

On a ici que les pages A et C sont d'importance égale, et que B est moins importante. Note que ce n'est pas la même chose que la ligue. C'est raisonnable que les deux approches donnerait pas exactement la même importance. Dans la ligue, si C aurait perdu contre A 15 fois, on aurait peut-être changé notre opinion sur l'ordre des équipes. Mais qu'une page web fait 15 liens vers une autre page ne devrait pas compter 15 fois.

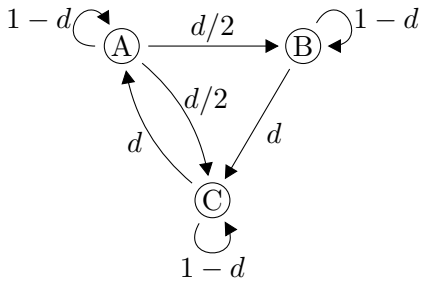
Exercice 5.6. Comment saviez-vous sans calcul que la valeur propre dominante est 1? □

On pourra aussi calculer des puissances de P afin de voir le vecteur d'état stationnaire dans les colonnes.

$$P^{10} = \begin{bmatrix} 0.40625 & 0.4375 & 0.37500 \\ 0.18750 & 0.1875 & 0.21875 \\ 0.40625 & 0.3750 & 0.40625 \end{bmatrix} \qquad P^{20} = \begin{bmatrix} 0.3994140625 & 0.400390625 & 0.4003906250 \\ 0.2001953125 & 0.199218750 & 0.2001953125 \\ 0.4003906250 & 0.400390625 & 0.3994140625 \end{bmatrix}$$

Exercice 5.7. Est-ce que $k = 10$ est "grand"? C'est-à-dire, est-ce que toutes les colonnes de P^{10} sont à peu près égales? Et $k = 20$? En vous fiant aux puissances données, donner une approximation du vecteur d'état stationnaire et la comparer au vecteur exact ci-haut. □

On peut faire mieux, en introduisant un peu de paresse: à chaque page web, on se permet l'option de ne rien faire, c'est-à-dire de rester sur la page présente. Posons que la probabilité de se déplacer est d. Cette probabilité se divise également entre les liens de la page, et le $1-d$ restant s'attache à la page présente. Voici le graphe et la matrice de transition Q.



$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 0,5 & 0 & 0 \\ 0,5 & 1 & 0 \end{bmatrix}$$

Les deux matrices de transition, P et Q, sont fortement reliés.

Exercice 5.8. Montrer que $Q = dP + (1-d)I$, où I représente la matrice d'identité. □

Exercice 5.9. Montrer que P et Q possèdent exactement les mêmes vecteurs propres (indice: calculer $I\mathbf{v}$ où $\mathbf{v}$ est vecteur propre de P). Quelles sont les valeurs propres de Q, en termes de celles de P? □

La conséquence est que le vecteur d'état stationnaire de P est exactement la même que le vecteur d'état stationnaire de Q. On pourra utiliser l'une ou l'autre matrice. La distinction, c'est que la paresse est plus *rapide*: une puissance de Q converge typiquement plus rapidement vers le vecteur d'état stationnaire qu'une puissance de P. Voyons:

$$Q^{10} \approx \begin{bmatrix} 0.3993 & 0.3997 & 0.4008 \\ 0.2004 & 0.1995 & 0.1998 \\ 0.4003 & 0.4008 & 0.3993 \end{bmatrix}$$

On voit clairement le vecteur d'état stationnaire, à trois décimales. À comparer avec P^{10}.

Il reste une autre optimisation (qui s'applique également aux chaînes de Markov et même au systèmes dynamiques).

Proposition 5.10. *Soit A une matrice. Les colonnes de A^k sont toutes approximativement multiples d'un même vecteur $\mathbf{v}$ si et seulement si $A^k\mathbf{x}_0$ est approximativement multiple de $\mathbf{v}$ pour tout vecteur $\mathbf{x}_0$.* □

Donc au lieu de calculer A^k, ou aurait pu calculer $A^k\mathbf{x}_0$ pour n'importe quel vecteur de départ $\mathbf{x}_0$. L'avantage c'est au plan technique: calculer A^k entraîne multiplier matrice par matrice k fois, tandis que $A^k\mathbf{x}_0$ ne requiert que multiplier matrice par vecteur k fois.

Exercice 5.11. Soit $\mathbf{x}_0 = \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \end{bmatrix}$. Calculer $P^{10}\mathbf{x}_0$, $P^{20}\mathbf{x}_0$ et $Q^{10}\mathbf{x}_0$ pour les matrices P et Q ci-haut (en utilisant les calculs de puissances déjà donnée). Vérifier qu'on obtient des approximations des vecteurs dominantes. Choisir d'autres vecteurs de départ et refaire.

Calculer à bras P^2, et ensuite calculer $P\mathbf{x}_0$, et $P(P\mathbf{x}_0)$. Expliquer pourquoi le calcul de $P^2\mathbf{x}_0$ et plus rapide que le calcul de P^2. □

Le calcul du vecteur d'état stationnaire de Q est essentiellement le calcul de PageRank de Google. Voir *The anatomy of a large-scale hypertextual search engine* sur l'internet...