

Analyse de réseaux multicouches pour une meilleure prédition du risque crédit[©]



María Óskarsdóttir^{a,*}, Cristián Bravo^b

^a Département d'informatique, Université de Reykjavík, Menntavegur 1, 102 Reykjavík, Islande

^b Département des sciences statistiques et actuarielles, Université de Western Ontario, 1151 Richmond Street, London, Ontario, N6A 3K7, Canada

Article info

Historique de l'article :

Reçu le 28 janvier 2021
Accepté le 13 juillet 2021
Disponible en ligne le 15 juillet 2021

Mots-clés :

Analyse commerciale
Risque de crédit
Science des réseaux
Réseaux multicouches Prêts agricoles

ABSTRACT

Nous présentons un modèle de réseau multicouche pour l'évaluation du risque de crédit. Notre modèle tient compte de multiples connexions entre les emprunteurs (telles que leur localisation géographique et leur activité économique) et permet de modéliser explicitement l'interaction entre les emprunteurs connectés. Nous développons un algorithme PageRank personnalisé multicouche qui permet de quantifier la force de l'exposition au défaut de tout emprunteur dans le réseau. Nous testons notre méthodologie dans un cadre de prêt agricole, où l'on a observé pendant longtemps des corrélations de défaut entre les emprunteurs lorsqu'ils sont soumis aux mêmes risques structurels. Nos résultats montrent qu'il y a des gains prédictifs significatifs rien qu'en incluant dans le modèle des informations sur la centralité des réseaux multicouches, et que ces gains sont accrus par des informations plus complexes telles que les variables PageRank multicouches. Les résultats suggèrent que le risque de défaillance est le plus élevé lorsqu'un individu est connecté à de nombreux défaillants, mais ce risque est atténué par la taille du voisinage de l'individu, ce qui montre que le risque de défaillance et la stabilité financière se propagent à travers le réseau.

2021 Elsevier Ltd. Tous droits réservés.

1. Introduction

La science des réseaux s'est imposée comme une théorie fondamentale et un outil essentiel pour comprendre, décrire, modéliser et analyser les systèmes complexes d'entités en interaction qui apparaissent naturellement, par exemple, en biologie, en sociologie, en finance et en économie, entre autres [4]. Dans la vie réelle, les objets sont souvent reliés par plus d'un type de relation. Par exemple, dans les réseaux sociaux en ligne, deux personnes peuvent être amies sur Facebook et se suivre sur Twitter. Ces réseaux sont généralement appelés réseaux multicouches, dont la théorie et le développement de modèles ont été au cœur de la science des réseaux au cours de la dernière décennie.

Dans le secteur bancaire, on soupçonne depuis longtemps l'existence de défaillances corrélées. Cependant, toutes les études portant sur la propagation du risque à travers les réseaux se sont concentrées sur la contagion entre les institutions financières [8,20] ou sur la corrélation entre les actifs et l'économie globale, proposée pour la première fois dans les accords de Bâle [5]. Plusieurs auteurs ont souligné que ces effets de corrélation affectent effectivement les mesures actuelles du risque de crédit [21,49] au niveau individuel, ce qui a été négligé jusqu'à présent. Compte tenu de l'importance des prêts aux particuliers pour l'ensemble de l'économie et des progrès actuels de la science des réseaux, nous combinons le risque de crédit avec la science des réseaux multicouches pour présenter une méthodologie permettant de calculer explicitement l'impact des défaillances corrélées.

[©] Domaine : Analyses basées sur les données. Ce manuscrit a été traité par l'éditeur associé Asil Oztekin.

* Auteur correspondant.

Adresse électronique : mariaoskars@ru.is (M. Óskarsdóttir).

dans les emprunteurs connectés à un niveau individuel. Notre proposition vise donc non seulement à mieux comprendre la propagation des défaillances, mais aussi à améliorer les systèmes d'évaluation des demandes, couramment utilisés pour prendre des décisions en matière d'octroi de prêts.

Le présent document comporte trois contributions principales. Premièrement, nous présentons un cadre pour la création d'un réseau bipartite multicouche à partir d'un ensemble de données tabulaires comportant au moins deux variables de connexion. Ces variables ne sont pas nécessairement des variables de réseau explicites, mais peuvent être une propriété partagée ou une caractéristique. Deuxièmement, nous développons une nouvelle approche pour la prédiction du risque de crédit avec une mesure de centralité PageRank multicouche personnalisée qui classe les nœuds du réseau multicouche par rapport à un ensemble de nœuds sources qui peuvent être utilisés pour comprendre le risque corrélé et également pour améliorer les systèmes de notation de crédit. Enfin,

nous montrons que les scores d'influence des réseaux multicouches contribuent à des modèles d'évaluation du crédit plus performants. Bien que nos résultats soient généraux et applicables à tout ensemble de données de prêt (et à tout problème dans lequel des effets corrélés sont suspectés), nous motivons notre travail dans un environnement dans lequel la défaillance corrélée est évidente. Nous concentrons notre attention sur les réseaux multicouches dans les prêts agricoles et présentons des algorithmes généraux pour les inclure dans tout modèle prédictif où ils apparaissent. Si les réseaux multicouches ont été étudiés dans le secteur bancaire, jusqu'à présent - et à notre connaissance - leurs effets sur les prêts de détail (prêts à la consommation traditionnels et prêts de gré à gré aux petites et moyennes entreprises) n'ont pas fait l'objet du même niveau d'attention. Cela n'est pas surprenant compte tenu de sa structure : les prêts de détail sont beaucoup plus nombreux, de sorte que toute interaction entre les réseaux nécessite une analyse complexe qui n'a été effectuée que dans le cadre de l'étude de l'effet des réseaux sur les prêts de détail.

sont devenus réalisables au cours de la dernière décennie. Cela dit, il est également naturel de s'attendre à ce que les emprunteurs soient connectés de nombreuses manières complexes, de sorte que l'analyse utilisée pour mesurer leur comportement devrait également suivre ce niveau de complexité.

Dans cet article, nous présentons une méthodologie pour la prédition du risque de crédit qui poursuit l'orientation de la recherche entamée par Óskarsdóttir et al [41] et Bravo et Óskarsdóttir [11]. Dans ces travaux antérieurs, il a été démontré que les informations provenant de réseaux sociaux à couche unique amélioraient la performance prédictive des modèles d'évaluation du crédit dans les prêts de détail, et qu'il existait une corrélation entre l'influence propagée par les réseaux multicouches bipartites et le comportement en cas de défaut, respectivement. Dans ce travail, nous construisons des réseaux multicouches bipartites, que nous étendons en incluant la force des interconnexions entre les couches, que nous considérons comme un *facteur d'adhérence*. Nous utilisons un ensemble de données réelles de prêts agricoles, un domaine d'application qui a été reconnu comme ayant de forts effets de réseau dans ses chaînes d'approvisionnement [6], et il a été démontré que les chaînes d'approvisionnement s'améliorent lorsqu'un crédit suffisant est disponible [55]. Nous relions les borrowers, c'est-à-dire les producteurs, au produit qu'ils fabriquent et au quartier dans lequel ils vivent pour créer des réseaux multicouches. Nous appliquons notre nouvel algorithme PageRank personnalisé pour les réseaux multicouches avec des mauvais payeurs connus comme source d'influence et extrayons ensuite des informations du réseau en termes de variables que nous ajoutons à l'ensemble de données sur les prêts. Ensuite, pour comprendre le pouvoir prédictif des variables dérivées de ces réseaux multicouches, nous les comparons à des modèles d'évaluation du crédit construits sans variables de réseau et avec des modèles à couche unique. Enfin, nous explorons les nouvelles connaissances issues de ces variables en utilisant les explanations additives de SHapley [SHAP ; 36] sur le modèle de renforcement du gradient stochastique le plus performant, afin d'interpréter les nouvelles variables et de comprendre les nouvelles connaissances qu'elles apportent à la pratique de l'évaluation du risque de crédit. Nos résultats montrent une amélioration significative des performances du modèle lorsque l'on inclut des variables de réseau multicouche dérivées du PageRank, ainsi qu'une interaction complexe et non linéaire entre les variables.

La structure de ce document est la suivante : la section suivante positionne ce document dans la littérature actuelle, à la fois dans la science des réseaux et dans le prêt au détail. La section suivante propose l'algorithme de Pagerank personnalisé multicouche qui sera la base de ce travail. La section 4 présente l'ensemble des données et les résultats expérimentaux. La dernière section présente la conclusion de l'article.

2. Travaux antérieurs

2.1. Centralité dans les réseaux multicouches

Dans les systèmes complexes, le type d'interactions entre les entités des différents sous-systèmes peut varier et la théorie des réseaux multicouches a donc évolué au sein de la science des réseaux au cours des dernières années pour modéliser ces interactions [4,30]. Outre les nœuds et les arêtes qui sont les éléments fondamentaux des réseaux ordinaires à une seule couche (ou graphes), les réseaux multicouches comportent plusieurs couches pour tenir compte des différents types d'interactions. Dans le cas le plus général des réseaux multicouches, tout nœud peut appartenir à n'importe quel sous-ensemble de couches et une arête relie deux nœuds quelconques dans n'importe quelle couche ou entre deux couches quelconques. Il existe différentes formes de réseaux multicouches, en fonction des caractéristiques des nœuds et des arêtes à l'intérieur et entre les différentes couches et des

contraintes auxquelles ils sont soumis. Par exemple, dans les réseaux multiplex, les arêtes inter-couches n'existent qu'entre le même nœud dans toutes les couches [30].

La centralité est un thème important dans la science des réseaux, où certains indicateurs de centralité sont utilisés pour identifier les nœuds les plus importants d'un réseau. Les mesures de centralité les plus courantes sont le degré, la proximité, l'interdépendance et la centralité du vecteur propre (ou PageRank). Le degré compte le nombre d'arêtes incidentes à un nœud et est à la fois conceptuel et informatiquement le plus simple. La proximité

La *Proximité* mesure la distance moyenne d'un nœud par rapport à tous les autres nœuds du réseau, et donc son degré de proximité par rapport aux autres nœuds. L'interdépendance représente la fréquence à laquelle un nœud se trouve sur le chemin le plus court entre les autres nœuds du réseau. La proximité et l'interdépendance sont coûteuses en termes de calcul, car elles nécessitent de trouver les chemins les plus courts entre toutes les paires de nœuds du réseau. La centralité du PageRank mesure l'influence d'un nœud dans un réseau. Elle dépend du nombre de nœuds liés au nœud et de leur centralité PageRank. Il a été conçu à l'origine pour classer les pages web dans les moteurs de recherche [44], mais a été utilisé dans de nombreuses applications [9,31,35,38]. Il est basé sur la distribution stationnaire d'un marcheur aléatoire dans un réseau qui peut se téléporter aléatoirement vers n'importe quel autre nœud au moyen d'un vecteur de redémarrage. Le vecteur de redémarrage de l'équation peut être manipulé pour orienter la marche aléatoire vers un ensemble de nœuds sources. C'est ce qu'on appelle le PageRank personnalisé [44]. Il a été utilisé dans diverses applications pour classer les nœuds en fonction de la source d'influence, de sorte que les nœuds les plus proches de la source obtiennent un score plus élevé [52].

Le développement de la recherche sur les réseaux multicouches s'accompagne de celui de la littérature sur la généralisation des mesures de centralité, en particulier la centralité de type PageRank. En raison des multiples types d'arêtes et de la complexité des couches supplémentaires, il ne s'agit pas d'une tâche directe. D'une part, plusieurs articles ont développé une mesure de centralité de type PageRank pour les réseaux multiplex. Halu et al [27], par exemple, ont étendu la mesure de centralité PageRank aux réseaux multiplexes en calculant un PageRank régulier dans une couche du réseau et en utilisant le classement obtenu comme entrée pour le classement de la deuxième couche. De même, Pedroche et al [46] ont généralisé la notation du PageRank classique dans le style des réseaux bplex et l'ont ensuite étendue aux réseaux multiplexes. Le PageRank multiplex fonctionnel est capable de saisir les effets non linéaires causés par différents types de liens entre les nœuds [28,29], tandis que Tu et al. [51] ont défini un PageRank multiplex de couplage qui peut décrire les effets de couplage donnés à chaque modèle distinct de connexions. D'autre part, une mesure de centralité PageRank multicouche a été développée dans MuxViz, un cadre pour l'analyse et la visualisation des réseaux multicouches [17]. Ce cadre comprend des extensions et une formulation mathématique de plusieurs concepts de réseau pour les réseaux multicouches, tels que la diffusion et la centralité [18,19,24,30]. Dans le cadre de MuxViz, le PageRank multicouche est calculé de la même manière que le PageRank normal, mais en utilisant des eigentors et en agrégeant les valeurs des nœuds qui apparaissent dans plus d'une couche. Cherian et al [16] ont étendu cette version de la mesure du PageRank multicouche en introduisant des poids pour les couches du réseau.

En résumé, la plupart des recherches menées jusqu'à présent ont porté sur la généralisation de l'algorithme PageRank pour les réseaux multicouches. L'une des contributions du présent document est un PageRank personnalisé pour les réseaux multicouches qui, à notre connaissance, n'a jamais été présenté auparavant.

2.2. Apprendre des réseaux

L'objectif de cet article est de prédire la solvabilité à l'aide d'informations sur les réseaux multicouches. En général, il s'agit d'un problème de classification binaire, où chaque observation - ou nœud - appartient à l'une des deux classes, en l'occurrence défaillant ou non défaillant. Les informations provenant des réseaux peuvent être obtenues de plusieurs manières. Tout d'abord, avec l'extraction de caractéristiques, des variables prédefinies décrivant le voisinage des nœuds et/ou les propriétés structurelles sont combinées pour les nœuds du réseau, ce qui augmente le nombre d'attributs des nœuds. Les attributs des nœuds peuvent ensuite être utilisés avec une technique

d'apprentissage supervisé pour prédire la solvabilité. Deuxièmement, l'apprentissage par représentation est une technique qui permet d'obtenir automatiquement une représentation des caractéristiques des nœuds du réseau. L'intégration peut ensuite être utilisée comme entrée pour une classification binaire.

fiers [25]. Enfin, les classes de nœuds peuvent être apprises directement à partir du réseau, à l'aide de techniques d'apprentissage en réseau [37] ou de techniques d'apprentissage en profondeur telles que les réseaux neuronaux graphiques.

Toutes les approches ont été utilisées avec succès pour les réseaux ordinaires à couche unique dans diverses applications. L'extraction de caractéristiques a été utilisée pour prédire le taux de désabonnement dans les télécommunications [42], pour la détection des fraudes à la sécurité sociale

[52] et l'évaluation du crédit [41]. L'apprentissage par représentation a été utilisé pour diverses tâches, notamment la prédiction de liens [25] et l'étiquetage des nœuds, comme la prédiction du désabonnement [43]. L'apprentissage en réseau avec des approches d'apprentissage profond est actuellement en plein essor [13,56].

Toutefois, la littérature ne s'étend pas aux réseaux multicouches en dehors de l'apprentissage de la représentation. L'apprentissage de la représentation est apparu récemment comme une technique permettant de cartographier les sommets et les informations concernant leurs caractéristiques et leurs propriétés structurelles dans un espace vectoriel de dimension inférieure [53]. En conséquence, le réseau est facile à manipuler pour les analyses ultérieures, telles que la prédiction des liens, la classification des nœuds et la détection des communautés. Au début, les efforts de recherche étaient principalement consacrés aux réseaux à une seule couche, mais au cours des deux dernières années, l'accent a été mis sur le développement de l'apprentissage de la représentation pour les réseaux multiplexes [2,14,34,40,54]. Bien que ces techniques soient capables de préserver à la fois la structure et le contenu des nœuds, et de s'adapter à de grands réseaux, leur principal inconvénient, en particulier dans le contexte du prêt, est le manque d'interprétabilité des encastrements appris.

Certaines recherches sur l'extraction des caractéristiques des réseaux multicouches ont été menées à l'étranger. Par exemple, Amoroso et al. [1] ont utilisé des réseaux multiplex sur des images IRM pour détecter les signes précoce de la maladie d'Alzheimer. Ils ont caractérisé les réseaux et utilisé des forêts aléatoires pour prédire les signes de la maladie. Dans notre approche, nous extrayons de la même manière les caractéristiques des réseaux multicouches et étudions leur performance prédictive en termes de risque de crédit.

Nous pouvons constater que la plupart des recherches existantes sur l'apprentissage à partir de réseaux utilisent des réseaux monocouches et unipartites. Nous extrayons des caractéristiques que nous concevons sur la base de réseaux multicouches et des sorties de l'algorithme de propagation.

2.3. Effets de réseau dans la mesure du risque de crédit de détail

Comme nous l'avons mentionné précédemment, il a été démontré que les connexions de réseau affectent le risque de crédit. Les connexions entre les banques ont été liées au risque systémique d'une économie [50]. Cette étude s'est toutefois concentrée sur les liens entre les entreprises qui participent à un système financier, et non sur l'emprunteur lui-même, qui est le client de ces entités. Une analyse similaire, mais à plusieurs niveaux, a été menée par Montagna et al [39], qui ont identifié que les risques de contagion des défaillances entre les entreprises financières ne sont pas linéaires. Enfin, très récemment, Gupta et Kumar [26] ont étudié l'utilisation de méthodes de regroupement pour identifier les effets de réseau bipartites dans les banques en Inde. Tous ces travaux suggèrent fortement que les effets de réseau peuvent accroître le risque du système dans son ensemble.

Du côté des emprunteurs, nos travaux antérieurs [41] ont montré comment la connectivité entre les emprunteurs au sein d'un même réseau affecte le risque de crédit. Les variables de réseau étaient parmi les prédicteurs les plus forts du défaut sur un ensemble de données de cartes de crédit. Nos travaux antérieurs ne se sont

Conscients de la complexité de l'évaluation du risque de crédit, Bai et al. [3] développent une méthode avec des ensembles flous dans le domaine du prêt agricole, afin d'extraire et d'étudier les relations complexes entre les caractéristiques des agriculteurs, les facteurs environnementaux compétitifs et la solvabilité des agriculteurs.

En conclusion, la littérature antérieure sur les effets de réseau sur le risque de crédit se concentre soit sur les institutions financières qui composent le marché, par opposition au niveau de l'emprunteur, soit sur un réseau unique au lieu d'un réseau multicouche. Dans ce travail, nous nous attaquons au problème plus complexe et plus général de la mesure du risque à travers des réseaux multicouches au niveau de l'emprunteur, puis de la construction d'un modèle prédictif à partir des résultats de l'analyse du réseau.

3. Centralité du PageRank multicouche personnalisée

3.1. Construction d'un réseau multicouche bipartite

Soit $G = (V, E)$ un réseau à une seule couche - un graphe - où V est un ensemble de nœuds et $E \subseteq V \times V$ est un ensemble d'arêtes qui relient des paires de nœuds. Un réseau est représenté par sa matrice d'adjacence ou son tenseur d'adjacence de rang 2 W^t qui encode des informations sur les connexions entre les nœuds.

Le réseau est constitué d'une série d'arêtes qui vont du nœud i au nœud j et dont l'intensité est exprimée en termes de poids de l'arête.

Si le réseau compte $N = |V|$ nœuds, W^t a une dimension $N \times N$. Dans un réseau bipartite $G = (V, V, E)$, chaque nœud appartient à l'une des catégories suivantes

basés sur la technologie de l'information et de la communication. ¹ ²
toutefois pas attardés sur les effets multicouches, qui se sont avérés pertinents pour estimer le risque. Dans [47], les auteurs ont souligné que les effets systémiques
les estimations du risque peuvent être gravement sous-estimées si les effets multicouches

de deux ensembles indépendants et disjoints, V_1 ou V_2 , et les arêtes existantes.

Le réseau bipartite n'existe qu'entre des nœuds appartenant à des ensembles différents, c'est-à-dire $E \subseteq V_1 \times V_2$. Soit $N_1 = |V_1|$ et $N_2 = |V_2|$. La matrice d'adjacence d'un réseau bipartite est donnée par la matrice $(N_1 + N_2) \times (N_1 + N_2)$ où les entrées des premiers blocs $N_1 \times N_1$ et des derniers blocs $N_2 \times N_2$ sont nulles puisqu'il n'existe pas d'arêtes entre les nœuds d'un même ensemble.

Les réseaux comportant plus d'un type d'arêtes sont représentés à l'aide de plusieurs couches où un type d'arêtes relie les nœuds de chaque couche. Ces réseaux sont appelés *réseaux multicouches*. Les nœuds de chaque couche sont connectés sur la base d'un type de relation, de sorte que les différentes couches sont utilisées pour tenir compte de tous les différents types de relations. Les arêtes peuvent exister entre n'importe quelle paire de nœuds dans la même couche, ou entre n'importe quelle paire de nœuds dans n'importe quelle paire de couches. On parle alors d'arêtes intra et inter, respectivement.

Un réseau multicouche M avec N nœuds et L couches peut être représenté par un réseau d'adjacence de rang 4 de dimension $N \times N$

ne sont pas prises en compte. Cette dernière étude s'est à nouveau concentrée sur un institut financier.

Le réseau de l'emprunteur n'est pas un réseau régional mais un réseau au niveau de l'emprunteur.

Une étude plus récente de Cheng et al. [15] se concentre directement sur les prêts en réseau. Ces prêts sont accordés non pas à un individu, mais à une clique dans laquelle tous les utilisateurs se portent garants les uns des autres. Leurs résultats suggèrent, sans surprise, que la composition de la clique est extrêmement importante pour le risque de défaillance de chaque individu. Les effets de ce travail sont toutefois basés sur le réseau explicite de garants, sans aller plus loin dans l'étude d'autres types de connexions.

$\times L \times L M^{\alpha}$ qui indique le poids d'une arête entre le nœud i dans le réseau M et le nœud j dans le réseau M^{α} .

couche α et le nœud j dans la couche β [18]. Par convention, nous représentons les nœuds par des lettres latines et les couches par des lettres grecques. Pour simplifier les notations et les calculs, le tenseur d'adjacence de rang 4 peut être aplati pour obtenir une représentation équivalente du réseau sous la forme d'un tenseur de rang 2 qui est une matrice $N - L \times N - L$, connue sous le nom de tenseur ou matrice d'adjacence supra.

Dans le présent document, nous considérons des réseaux multicouches avec un réseau bipartite dans chaque couche. Nous supposons qu'il existe un ensemble de nœuds présents dans toutes les couches, appelés *nœuds communs*, désignés par V_c , le nombre de nœuds communs étant égal à $|V_c|$. Nous supposons en outre que chaque couche possède son propre ensemble de *nœuds spécifiques* qui n'existent que dans cette couche spécifique. Étant donné L couches $\alpha, \beta, \dots, \omega$, l'ensemble des nœuds spécifiques de chaque couche est $V^{\alpha}, V^{\beta} \dots V^{\omega}$, avec le nombre de nœuds spécifiques dans chaque couche est donné par $|V^{\alpha}|, |V^{\beta}|, \dots, |V^{\omega}|$.

Lors de la construction de la matrice d'adjacence supra de ces types de réseaux, nous considérons que $N = |V_c| + |V^{\alpha}| + |V^{\beta}| + \dots + |V^{\omega}|$ est le nombre de réseaux qui ont été mis en place dans le cadre de l'initiative de la Commission européenne.

de nœuds distincts dans le réseau. Cela signifie que tous les nœuds spécifiques existent dans chaque couche, mais seules les arêtes entre les nœuds communs et les nœuds spécifiques de la couche sont autorisées. Cela permet de s'assurer que les dimensions des matrices correspondent. Dans ces réseaux multicouches, les arêtes intercouches n'existent qu'entre les nœuds communs, c'est-à-dire les nœuds noirs de la figure 1, où les arêtes intercouches sont représentées par des lignes en pointillés. Le poids des arêtes inter-couche donne une indication de l'efficacité du réseau.

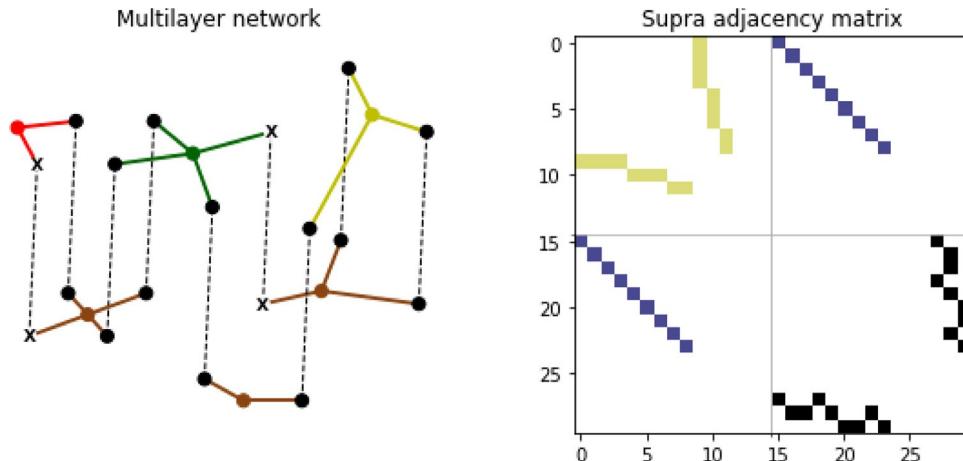


Fig. 1. Un réseau multicouche (à gauche) et sa matrice d'adjacence supra (à droite). Le réseau multicouche comporte neuf nœuds d'emprunteur (noir), trois nœuds de localisation (marron) et trois nœuds de produit (rouge, vert, jaune) en deux couches. Les nœuds de la première couche sont reliés par des arêtes brunes et les nœuds de la deuxième couche sont reliés par des arêtes rouges, vertes et jaunes. Les arêtes intermédiaires (gris foncé) relient un nœud emprunteur à lui-même dans l'autre couche. Dans la matrice d'adjacence supra, la sous-matrice supérieure gauche (jaune) indique la matrice d'adjacence pour la première couche, la matrice inférieure droite (noire) est la matrice d'adjacence pour la deuxième couche. Les sous-matrices supérieure droite et inférieure gauche (bleues) sont les matrices d'adjacence pour les inter-arrêtes. Elles sont diagonales car chaque nœud n'est lié qu'à lui-même dans l'autre couche. Les nœuds noirs marqués d'un X sont la source d'influence. (Pour l'interprétation des références aux couleurs dans la légende de cette figure, le lecteur est invité à se reporter à la version web de cet article).

Il s'agit d'une mesure de la facilité avec laquelle il est possible de passer d'une couche à l'autre, un poids plus élevé indiquant une transition plus probable. Nous appelons ce paramètre de poids inter-couches "stickiness" (*adhérence*), S . La valeur par défaut de S est 1, ce qui signifie qu'il n'est pas pondéré. Notez que lorsque $S \rightarrow 0$, la connexion entre les couches devient de plus en plus faible jusqu'à ce que les couches finissent par être déconnectées. En revanche, lorsque $S \rightarrow \infty$, les couches sont agrégées, ce qui donne un réseau avec une seule couche mais plusieurs types d'arêtes, c'est-à-dire un réseau dont les arêtes sont colorées.

La sous-figure de gauche de la [figure 1](#) montre un exemple d'un tel réseau multicouche à deux couches. Il y a neuf nœuds communs (noir), trois nœuds qui appartiennent uniquement à la couche inférieure (marron) et trois nœuds qui appartiennent uniquement à la couche supérieure (rouge, vert ou jaune).

Pour faciliter la notation, nous supposons, sans perte de généralité, qu'il existe deux couches, α et β , avec des matrices d'adjacence intra-couche

A^α et B^β , les nœuds communs V_c , et les nœuds spécifiques V_α et V_β dans les couches α et β respectivement. La matrice d'adjacence supra

$M_{\alpha\beta}^{\alpha\beta}$ peut s'écrire comme suit

$$M_{\alpha\beta}^{\alpha\beta} = \begin{array}{c|cc} & \begin{matrix} A^\alpha & I^\alpha \\ \hline I^\beta & B^\beta \end{matrix} & \end{array}$$

où les matrices A , B , I sont de dimension $N \times N$ avec $N = |V_c| +$

$\alpha \quad \beta$

$|V_\alpha| + |V_\beta|$ et I est nul à l'exception des valeurs sur la diagonale corrélatives à celles de l'indice.

Le tenseur d'adjacence aplati du réseau multicouche de gauche est représenté dans la sous-figure de droite. La sous-figure de droite de la [figure 1](#) montre le tenseur d'adjacence aplati pour le réseau multicouche de gauche. Cette méthode peut facilement être étendue à plus de deux couches, en ajoutant des matrices d'adjacence intra-couche à la diagonale de la matrice d'adjacence supra et des matrices d'adjacence inter-couche à la diagonale opposée, tout en conservant la symétrie.

Pour décrire les probabilités de transition d'un marcheur aléatoire traversant le réseau multicouche, à la fois à l'intérieur et

un autre nœud au hasard dans le réseau. Dans un réseau à une seule couche, il s'agit de la solution en régime permanent de l'équation $p_j(t+1) = R^t p_i(t)$

où R^t est la matrice de transition d'un marcheur aléatoire qui se rend à un nœud voisin avec une probabilité r mais saute à tout autre nœud du réseau avec une probabilité de $1 - r$. Le paramètre r est appelé facteur de redémarrage ou d'amortissement. Il détermine le compromis entre l'importance du réseau et le saut aléatoire dans la marche aléatoire. Traditionnellement, il est fixé à 0,85, car cette valeur garantit une convergence rapide tout en représentant avec précision le comportement de la navigation sur Internet : un internaute clique sur les hyperliens environ cinq sixièmes du temps, mais visite une nouvelle page un sixième du temps [7,12,32].

Dans un réseau multicouche, la marche aléatoire peut se déplacer entre les nœuds d'une même couche ainsi qu'entre les couches le long des arêtes intercouches. De même, le marcheur peut sauter à n'importe quel nœud de n'importe quelle couche. Par conséquent, le tenseur de transition de la marche aléatoire devient

$$R^t \alpha = R^t \alpha + \frac{1-r}{N-L} u^\alpha, \quad (1)$$

$$\beta \quad \beta \quad \frac{r}{N-L} \beta$$

où $T_{\alpha\beta}^t$ est le tenseur de transition supra de rang 2 du réseau, r est le facteur d'amortissement et u^α est une matrice $N - L \times N - L$ de 1, à indiquer une probabilité égale de passer à n'importe quel nœud de la couche multiple réseau.

Le tenseur de transition R^t représente la marche aléatoire, c'est-à-dire que le entre les couches, nous calculons la matrice de transition supra T^t , c'est-à-dire la colonne-matrice d'adjacence normalisée [22].

3.2. Centralité du PageRank multicouche

La centralité de PageRank d'un nœud équivaut à la probabilité qu'un marcheur aléatoire traversant le réseau aboutisse à ce nœud. En plus de se déplacer le long des arêtes du réseau, le marcheur aléatoire peut également sauter - ou se téléporter - vers un nœud donné.

la probabilité de passer d'une paire de nœuds à l'autre. Soit $p_{ia}(t)$ une tenseur dépendant du temps qui représente la probabilité de trouver le marcheur dans un nœud et une couche donnés à l'instant t . La marche aléatoire est donc

$$pj\beta(t+1) = R^i \alpha p_i \alpha(t).$$

La centralité PageRank multicouche est la solution à l'état stable de cette équation, c'est-à-dire lorsque $t \rightarrow \infty$, et est donnée par π_{ia} , qui représente la probabilité de trouver le marcheur au nœud i dans la couche α . La solution est obtenue en trouvant l'eigentor principal, ce qui est équivalent à la solution du problème des valeurs propres d'ordre supérieur,

$$T^{ia} \pi_{ia} = \lambda \pi_{ia}$$

comme indiqué dans [19].

La solution de l'état stable π_{ia} donne la probabilité de trouver le marcheur au nœud i dans la couche α . Pour atteindre finalement une valeur pour la centralité du PageRank pour chaque nœud, les valeurs dans toutes les couches sont ag-.

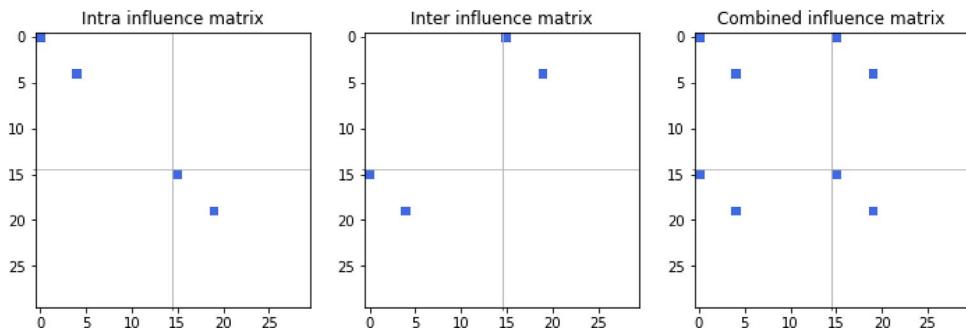


Fig. 2. Trois scénarios pour la matrice d'influence d'un réseau multiplex à deux couches lors du calcul d'un PageRank personnalisé.

La centralité du PageRank multicouche du nœud i est donc ω_i : la centralité du PageRank multicouche du nœud i .

3.3. Matrice d'influence

La centralité PageRank multicouche dérivée ci-dessus donne une représentation de l'importance des nœuds, en supposant que la marche aléatoire saute à n'importe quel autre nœud dans n'importe quelle couche avec une probabilité égale. Cependant, dans certaines situations, il est utile de faire ressortir les nœuds qui sont importants du point de vue d'un ensemble de nœuds spécifiques. Pour ce faire, on autorise la marche à sauter uniquement vers un ensemble de nœuds spécifiques, ce qui rend la marche aléatoire biaisée. Cette version est appelée PageRank personnalisé [44] et a par exemple été utilisée pour détecter les fraudes à la sécurité sociale [52].

Bien que le PageRank ait été dérivé pour les réseaux multicouches, à notre connaissance, le PageRank personnalisé multicouche a été dérivé pour la première fois par Bravo et Óskarsdóttir [11]. Nous décrivons ici comment le PageRank multicouche peut être généralisé pour les marches aléatoires biaisées afin d'obtenir un score personnalisé.

En partant de l'[Eq. \(1\)](#), nous modifions la matrice $u^i \alpha_j$, en fixant

des valeurs spécifiques égales à un et les autres valeurs à zéro, ce qui permet à la marche aléatoire de ne sauter que vers les nœuds d'un ensemble spécifique, que nous appelons nœuds d'influence ou source d'influence, désignés par $V_I \subset V$. Nous appelons la matrice $u^i \alpha_j$ la matrice d'influence. Comment...

Cependant, la mise à zéro de toutes les valeurs de la matrice d'influence, à l'exception de celles qui sont à l'origine de l'influence, n'est pas triviale, car ces nœuds apparaissent plusieurs fois dans $u^i \alpha_j$, à la fois à l'intérieur des couches et entre les couches. Pour il-

Pour illustrer cela, supposons qu'il y ait deux couches, de sorte que u ait une dimension de $2N \times 2N$ et puisse être divisé en quatre blocs de sous-matrices $N \times N$. Deux de ces blocs sont sur la diagonale, u_{11} et u_{22} , et correspondent aux bords intra-couche. Deux de ces blocs sont sur la diagonale, u_{11} et u_{22} , et correspondent aux bords intra-couche ; les deux autres blocs sont hors diagonale, dans le triangle supérieur et inférieur, et correspondent aux bords inter-couche. Nous les appelons u_{12} et u_{21} . La matrice se présente alors comme suit

$$u^i \alpha_j = \begin{array}{c|c} u_{11} & u_{12} \\ \hline u_{21} & u_{22} \end{array}$$

Nous proposons trois scénarios pour définir la matrice d'influence.

1. Intra-influence : Dans ce cas, la marche aléatoire passe par des nœuds V_I à l'intérieur des couches. Nous attribuons la valeur 1

aux nœuds V_I situés sur la diagonale des matrices intra-couche u_{11} et u_{22} .

2. Inter-influence : Dans ce scénario, l'influence provient d'entre les couches. Nous attribuons la valeur 1 aux nœuds V_I sur la diagonale des deux matrices inter-couche, u_{12} et u_{21} .
3. Influence combinée : Dans ce cas, l'influence provient à la fois de l'intérieur et de l'extérieur de l'entreprise. entre les couches, nous attribuons donc la valeur 1 aux nœuds V_I sur la diagonale des quatre sous-matrices de $u^i \alpha_j$.

Enfin, le dénominateur du second terme de l'équation (1) se réduit à la somme des éléments de $u^i \alpha$. Avec une transition redéfinie dix-^j pour la marche aléatoire, nous procédons au calcul de la centralité du PageRank comme dans la section 3.2.

β

4. Résultats expérimentaux

4.1. Ensemble de données

Pour ce travail, nous utilisons l'ensemble de données présenté précédemment dans [10]. Les statistiques descriptives des données et les descriptions des variables figurent à l'annexe B. Cet ensemble de données se concentre sur les prêts agricoles dans un pays d'Amérique latine, avec un ensemble de plus de 70 000 prêts d'une durée de un à cinq ans, couvrant 15 années de données (1998-2013). L'ensemble de données comprend des informations relatives au prêt, telles que le montant, la durée, l'existence de garanties et/ou de sûretés, et la valeur de ces dernières le cas échéant. Il comprend également des informations sociodémographiques sur l'emprunteur, notamment son âge, le district dans lequel il opère, le(s) produit(s) qu'il commercialise et, le cas échéant, toute information concernant des prêts antérieurs (nombre, montants, durées, arriérés antérieurs et autres comportements financiers antérieurs). Les données ont été préalablement nettoyées, en supprimant uniquement les valeurs extrêmes qui ne seraient pas conformes à la politique de l'institution financière, ainsi que les valeurs nulles.

4.2. Mise en place du réseau, calcul des scores et extraction des variables

L'objectif de notre étude de cas est de compléter les informations spécifiques aux prêts contenues dans l'ensemble de données par des informations sur les réseaux et, plus particulièrement, de quantifier l'influence des prêts en défaut de paiement. Pour ce faire, nous construisons des réseaux multicouches auxquels nous appliquons l'algorithme PageRank personnalisé afin d'obtenir des scores d'influence ainsi que d'autres caractéristiques de réseau. Ces caractéristiques sont ensuite ajoutées à l'ensemble des données relatives aux prêts, avant de construire des modèles d'évaluation du crédit pour les emprunteurs. Nous décrivons ci-dessous le processus en détail.

Tout d'abord, nous utilisons les données pour construire une séquence de réseaux multicouches comme suit. Chaque réseau multicouche est créé à partir des prêts accordés au cours d'une période de cinq ans.¹ Chaque réseau se compose de deux couches. Le choix des couches de réseau à utiliser est une étape très importante de la modélisation multicouche. Les réseaux choisis doivent être ceux qui, selon le modélisateur, propagent l'effet étudié (défaut dans ce cas) pour les individus étudiés. Dans de nombreuses circonstances, les réseaux disponibles seront simplement "ceux qui sont disponibles", mais dans d'autres circonstances, il peut y avoir des réseaux qui ne sont pas disponibles.

¹ Cette période a permis d'obtenir des réseaux dont presque tous les nœuds (> 99,5 %) se trouvent dans la même composante connectée, ce qui est essentiel pour obtenir des résultats stables de l'algorithme PageRank à l'étape suivante.

Tableau 1

Valeurs de l'AUC pour différentes valeurs du paramètre r sur l'ensemble d'évaluation, toutes variables confondues.

r	AUC (ensemble de validation)
0.2	0.733
0.3	0.737
0.4	0.734
0.5	0.734
0.6	0.734
0.7	0.738
0.8	0.738
0.85	0.738
0.95	0.735

de nombreux choix. Rien n'empêche le modélisateur potentiel d'inclure un grand nombre de réseaux, mais il doit toujours garder à l'esprit que la complexité des calculs croît avec le carré de la taille du réseau, comme indiqué ci-dessous.

Pour nos expériences, nous utiliserons le district géographique et le produit comme les deux couches de notre modèle. Ces deux facteurs ont été reconnus comme étant parmi les plus importants dans le domaine des prêts agricoles, en particulier dans le segment des petits agriculteurs [45]. La première couche que nous utilisons est celle du produit, où chaque emprunteur est connecté à son produit respectif, de sorte que plusieurs emprunteurs partagent une connexion s'ils plantent le même produit. Il est à noter que les agriculteurs peuvent planter plus d'un produit par an (rotation des cultures), un agriculteur peut être connecté à plus d'un autre agriculteur en utilisant ce réseau de produits. La deuxième couche est celle du district, où chaque borrhower est connecté à son district respectif et à sa zone respective (plusieurs zones constituent un district). En fait, nous avons fixé le paramètre d'adhérence entre les réseaux de zone et de district à ∞ , agrégant ainsi les deux couches. Nous avons procédé ainsi car les zones sont des sous-ensembles de districts et sont donc très étroitement liées. Par conséquent, cela signifie que les emprunteurs qui vivent dans la même zone sont plus étroitement liés que ceux qui vivent uniquement dans le même district. Chacune des deux couches consiste donc en un réseau bipartite. Pour compléter la construction du réseau, nous connectons chaque emprunteur avec lui-même dans les deux couches. Ces arêtes inter-couches peuvent être affectées de poids pour représenter le coût du déplacement entre les deux couches. Cette adhérence est l'un des paramètres que nous ajustons à la section 4.3.

Ensuite, nous appliquons l'algorithme PageRank personnalisé multicouche au réseau afin de calculer les scores d'influence. L'objectif est d'évaluer l'influence des anciens emprunteurs sur les nouveaux. Pour ce faire, nous incluons dans \mathcal{V} tous les emprunteurs qui ont fait défaut au cours des quatre premières années et onze mois de la période de cinq ans. Ils sont la source d'influence dans la matrice d'influence. Nous appliquons ensuite l'algorithme et inspectons les scores obtenus par les emprunteurs au cours du dernier mois de la période de cinq ans. Ces emprunteurs et leurs scores sont ensuite utilisés pour construire les modèles d'évaluation du crédit.

Les scores PageRank personnalisés dépendent du paramètre de redémarrage r . Il contrôle le compromis entre l'effet du réseau et l'influence des nœuds sources lors du calcul du score. r peut prendre n'importe quelle valeur entre 0 et 1, c'est pourquoi le réglage de ce paramètre est inclus dans nos expériences.

Outre les scores PageRank personnalisés, nous extrayons d'autres informations concernant les emprunteurs apparaissant dans le dernier mois des réseaux multicouches sur cinq ans, voir le tableau A.3 de l'annexe A. Inspirés par Getoor [23], nous comptons le nombre d'emprunteurs et le nombre d'emprunteurs défaillants, au cours de l'année et des cinq années précédentes, dans le voisinage le plus

proche de chaque nœud dans les deux couches du réseau et dans leur intersection. Nous obtenons ainsi les variables de degré du tableau A.3. En outre, nous extrayons le score du produit/district/zone auquel un emprunteur est connecté, c'est-à-dire le nœud spécifique dans chaque couche. Si un emprunteur est connecté à plus de

Un noeud spécifique dans la même couche, nous ne prenons en compte que la valeur la plus élevée.

Pour chaque valeur de r et de S , nous exécutons l'algorithme PageRank personnalisé multicouche trois fois : une fois avec la matrice d'influence intra, une fois avec la matrice d'influence inter et une fois avec la matrice d'influence combinée. Au terme de ce processus, les emprunteurs acquièrent un ensemble de variables de réseau, qui sont énumérées dans le [tableau A.3](#) de l'annexe A.

En outre, nous agrégeons les couches et considérons le cas où $S \rightarrow \infty$, c'est-à-dire lorsqu'il n'y a qu'une seule couche, mais deux types d'arêtes : le produit et le district. Nous calculons le PageRank personnalisé régulier avec des mauvais payeurs connus comme source d'influence pour différentes valeurs de r . Cette approche représente la manière naïve d'appliquer le PageRank personnalisé au réseau d'emprunteurs.

Pour obtenir les variables de réseau pour tous les emprunteurs de l'ensemble des données, nous répétons ce processus, en construisant à chaque étape un réseau s'étendant sur cinq ans et en calculant les scores des "nouveaux" emprunteurs, c'est-à-dire ceux du dernier mois, puis en décalant la période de cinq ans d'un mois, ce qui nous permet d'examiner un nouvel ensemble d'emprunteurs. Cela nous permet d'évaluer l'influence des mauvais payeurs antérieurs sur les nouveaux emprunteurs en ajoutant leurs variables de réseau à l'ensemble des données. Il convient de noter que les emprunteurs qui ont obtenu leur prêt au cours des quatre premières années et onze premiers mois de notre période d'observation doivent être exclus des exercices d'évaluation du crédit, car nous ne disposons pas d'informations suffisantes sur les mauvais payeurs antérieurs pour eux.

4.3. Mise en place de modèles prédictifs et réglage des paramètres

Avant d'exécuter le modèle, nous avons d'abord supprimé les variables fortement corrélées, en fixant le seuil à 70 % de corrélation, étant donné que la distribution des corrélations était en forme de U, avec un groupe dont la corrélation était inférieure à 30 %, presque aucune dans la fourchette 30-70 %, et le reste au-dessus du seuil de 70 %. En général, en cas de choix entre des variables corrélées, nous avons choisi la plus corrélée avec le target. Sur l'ensemble initial de 32 variables de réseau (voir le [tableau A.3](#) de l'annexe A), seules sept ne sont pas corrélées avec d'autres variables. En particulier, la plupart des variables ProdDegree sont corrélées avec les variables ProdAreaDegree combinées, à l'exception de ProdDegree5, la centralité à long terme associée au réseau de produits. Toutes les variables de district sont, comme prévu, corrélées avec les variables de zone, mais ces dernières sont plus informatives (mesurées par leur corrélation avec la variable par défaut) et nous ne les conservons donc que dans le modèle. Un autre résultat intéressant est qu'il existe une forte corrélation entre presque toutes les variables de réseau d'un an et leurs homologues de cinq ans, mais les variables d'un an sont mieux corrélées avec la variable par défaut (à l'exception de la centralité du réseau de produits, ProdDegree5). Cela signifie que, pour cet ensemble de données particulier, les effets à court terme sont plus importants que les effets à long terme. Ce phénomène ne se reproduira pas nécessairement dans d'autres ensembles de données, c'est pourquoi nous recommandons aux utilisateurs d'effectuer la même analyse que celle que nous venons de réaliser. Enfin, lors de la sélection des versions de la matrice d'influence à utiliser, nous avons constaté qu'elles avaient tendance à être très corrélées, et nous avons donc décidé d'utiliser la variable "combinée", puisqu'elle contient les informations de la combinaison des versions inter et intra, plus l'effet multicouche combiné. Cela signifie que l'une des dernières variables est redondante. La variable survivante, cependant, est fortement corrélée avec le score de pagerank du réseau agrégé (Aggregate). Cela signifie que seuls les scores multicouches et les scores aplatis restent dans l'ensemble de données, car ils sont les plus informatifs de l'ensemble. La [figure 3](#) montre les matrices corrélées finales, en différenciant les variables réseau et non réseau.

Notez qu'il n'y a pas de corrélation entre les variables des deux groupes.

Nous suivons le processus en comparant deux modèles différents. La régression logistique est la norme du secteur, représentant plus de 95 % de tous les modèles utilisés dans les banques [48], en raison de sa simplicité et de sa transparence. D'un autre côté, XGBoosting a été

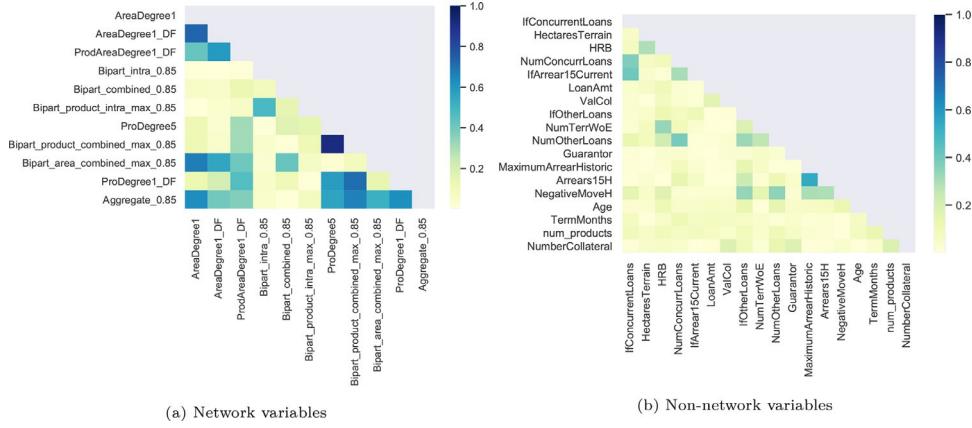


Fig. 3. Matrices de corrélation pour les deux ensembles finaux de variables.

s'est avéré être un ensemble d'arbres puissant, atteignant des performances de pointe dans la plupart des ensembles de données structurées sur le risque de crédit [33].

Pour le modèle de régression logistique, nous utilisons la pénalisation LASSO afin de réduire le nombre de variables sans signification, ce qui est particulièrement important étant donné que nous évaluons les variables du réseau. Nous ajustons le paramètre λ (le poids de pénalisation) en utilisant une procédure de validation croisée à trois niveaux sur un sous-ensemble de 20 % de la base de données de formation.

Le XGBoosting, quant à lui, dispose d'un ensemble de paramètres plus important à régler. Le XGBoosting, quant à lui, a un plus grand nombre de paramètres à régler. Nous réglons, pour chaque sous-ensemble de variables, le nombre d'arbres ($n_estimators \in [50, 100, 250, 500]$) en utilisant la notation `xgboost` bien connue du paquetage Python/R), la profondeur maximale de chaque arbre ($max_depth \in [2, 3, 4]$), le taux d'apprentissage ($learning_rate \in [0.001, 0.01, 0.1]$) et le régulateur de régression ($reg_alpha \in [0.1, 0.2, \dots, 1]$). Nous utilisons à nouveau un sous-ensemble de 20 % des données d'apprentissage et une procédure de validation croisée triple. Chaque modèle, construit sur différents sous-ensembles de variables, utilise alors un ensemble (distinct) de paramètres optimaux dans le but d'isoler l'impact des variables elles-mêmes au lieu d'un simple choix de paramètres.

La dernière série d'optimisation des paramètres qui a été réalisée concernait les réseaux eux-mêmes. Comme nous l'avons vu à la section 3.2, deux paramètres affectent les variables du réseau : le paramètre r qui contrôle le compromis entre la matrice et la matrice d'influence u_B^α , et le poids des arêtes intercouches

qui détermine le coût du déplacement entre les couches, l'adhérence S . Les deux paramètres ont été réglés en sélectionnant la meilleure aire sous la courbe de la caractéristique d'exploitation du récepteur (AUC) pour le modèle avec toutes les variables. Comme le modèle avec toutes les variables contient la plus grande quantité d'informations et que la régression logistique régularisée LASSO et le modèle XGBoosting sont tous deux dotés de capacités de sélection des variables (régularisation), nous pensons que cela conduit aux meilleurs paramètres pour toutes les autres expériences, car elles s'exécutent dans des sous-ensembles de l'ensemble de données complet. L'adhérence a eu peu d'influence sur les résultats, avec une détérioration significative des valeurs AUC pour $S = 0,25$ et $S = 4$, mais aucun effet ailleurs dans l'ensemble de données. Nous recommandons donc d'utiliser une adhérence neutre (c'est-à-dire $S = 1$) dans la construction du réseau, ce qui suppose qu'une entité a autant de chances de rester dans son propre réseau que de passer à une autre couche. La valeur r a un comportement plus intéressant. Le tableau 1 montre

la valeur AUC sur l'ensemble de validation pour différentes valeurs de r . La conclusion est qu'une valeur élevée de r (c'est-à-dire un poids élevé sur la matrice d'adjacence) est nécessaire pour optimiser la performance des variables du réseau. Dans nos expériences, les valeurs de 0,8 et 0,85 donnent des résultats équivalents, et nous avons donc choisi la valeur de 0,85, proposée dans l'article PageRank original.

Une fois tous les paramètres définis, nous pouvons maintenant discuter de la performance du modèle pour différents ensembles de variables.

Le tableau 2 montre la performance des modèles mesurée en termes d'AUC. Les modèles ont été construits avec différentes combinaisons de variables. Comme les variables comportementales, lorsqu'elles sont disponibles, tendent à être les plus importantes dans le risque de crédit [10], nous testons indépendamment les ensembles de données sociodémographiques et de prêts (ensemble de variables de base) de l'ensemble de variables comportementales. Les variables de réseau sont divisées en fonction de leur complexité et de leur niveau d'origine. L'ensemble de centralité comporte quatre sous-ensembles : Area and Product (centralité de degré et de degré par défaut calculée uniquement sur la couche area et la couche product), single networks (les deux variables ajoutées au modèle, calculées sur les couches simples), et single + multiplex (la centralité simple plus une variable de centralité additionnelle calculée sur le réseau multicouche). Les mesures de centralité sont le degré et le degré de défaillance, c'est-à-dire le nombre de défaillants dans le quartier, au cours de l'année précédente ou des cinq années précédentes. Le deuxième groupe est constitué des variables PageRank personnalisées, calculées sur le réseau multicouche à l'aide du processus décrit à la section 3.2. Le dernier sous-ensemble comprend toutes les variables du groupe réseau. Au total, nous avons testé 14 sous-ensembles de variables différents dans les deux modèles de référence, soit 28 expériences. Toutes les valeurs de SSC rapportées proviennent d'un ensemble de test indépendant.

Les résultats montrent que les modèles incluant toutes les variables de réseau sont significativement meilleurs que les deux modèles de base (avec et sans comportement). Cette différence est réduite comme prévu pour les modèles incluant des variables comportementales (3,4 % d'amélioration pour le modèle XGBoosting de toutes les variables contre 11,7 % pour le modèle de base + tous les réseaux), et ce gain est systématiquement plus élevé pour le modèle XG- Boosting que pour le modèle de régression logistique, pour tous les sous-ensembles de variables. Cela suggère que les variables du réseau ont une forte capacité prédictive non linéaire.

Une autre conclusion intéressante du tableau 2 concerne l'utilité d'utiliser plusieurs réseaux pour calculer les mesures de centralité, même sans utiliser une procédure de propagation plus sophistiquée, telle que le PageRank personnalisé multicouche. La ligne de base

Le modèle de centralité + (réseaux uniques), qui inclut les variables de centralité extraites de chaque réseau indépendant, est amélioré d'un point de pourcentage par rapport au modèle incluant toutes les variables de réseau dans les benchmarks Baseline et Behaviour (11,1 % contre 11,7 % dans le modèle Baseline et 3,3 % contre 3,4 % dans le modèle Behaviour). Cela montre que l'utilisation de réseaux multiples peut apporter des avantages aux modélisateurs, mais que ceux qui cherchent à extraire chaque point de précision seront mieux servis par l'approche PageRank personnalisée multicouche plus sophistiquée.

Pour approfondir l'importance des attributs, nous avons utilisé la méthode TreeSHAP [36] pour calculer la valeur moyenne de Shapley des attributs suivants

Tableau 2

AUC pour les différents modèles, et différence par rapport au modèle de base respectif (avec ou sans variables comportementales). Le modèle le plus performant dans chaque groupe est en gras.

$r = 0,85, S = 1$	Régression logistique		XGBoosting	
	Test AUC	% d'augmentation	Test AUC	% d'augmentation
Base de référence	0.639	-	0.660	-
Base + Centralité (Produit)	0.656	2.7%	0.707	7.1%
Base + Centralité (zone)	0.695	8.8%	0.719	8.9%
Base + Centralité (réseaux uniques)	0.698	9.2%	0.733	11.1%
Base + Centralité (simple + multicouche)	0.698	9.2%	0.729	10.5%
Base + PageRank personnalisé	0.648	1.4%	0.695	5.3%
Base + toutes les variables du réseau	0.703	10.0%	0.737	11.7%
Base + comportement	0.788	-	0.799	-
Base + Comportement + Centralité (Produit)	0.794	0.8%	0.818	2.4%
Base + Comportement + Centralité (zone)	0.802	1.8%	0.818	2.4%
Base + Comportement + Centralité (réseau unique)	0.805	2.2%	0.825	3.3%
Base + Comportement + Centralité (tous)	0.805	2.2%	0.824	3.1%
Base + comportement + PageRank personnalisé	0.791	0.4%	0.814	1.9%
Toutes les variables, y compris le comportement	0.807	2.4%	0.826	3.4%

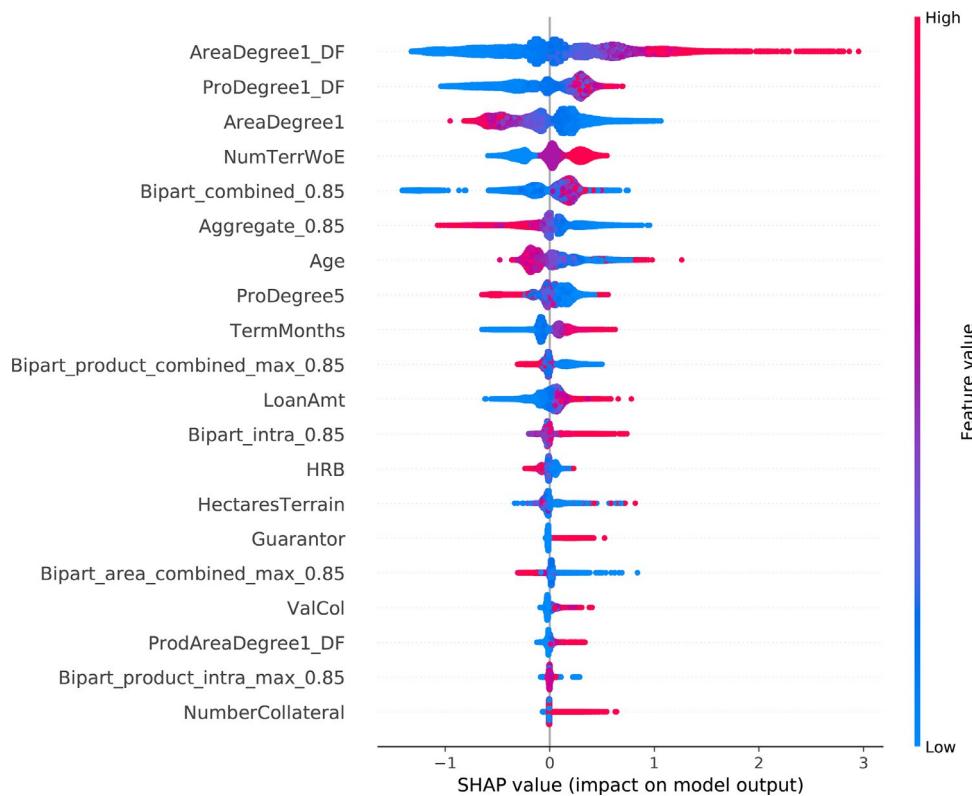


Fig. 4. Résumé de la valeur moyenne de Shapley pour un sous-ensemble de cas. Ordre d'importance des variables.

chaque variable sur un sous-ensemble de l'ensemble de données, pour le modèle de base + toutes les variables du réseau. Cela nous permet de créer la figure 4 qui montre les variables les plus prédictives (de haut en bas par ordre d'importance) mais aussi la valeur que prennent les variables, des valeurs les plus faibles en bleu aux valeurs les plus élevées en rouge. Nous pouvons également voir comment ces variables affectent la prédiction, puisqu'une valeur fortement positive (négative) sur l'axe horizontal est associée à un effet fortement positif (négatif) sur la probabilité de défaillance.

La toute première conclusion est que les variables de réseau dominent les caractéristiques les plus importantes. Les variables de centralité pour les emprunteurs défaillants (montrant à quel point les emprunteurs sont "proches" des cas de défaillance antérieurs) sont les variables les plus significatives du modèle. Il s'agit là d'une preuve

très forte de l'existence d'une corrélation entre les défaillances : plus un cas est lié à d'autres défaillants dans le réseau, plus il se traduit par une corrélation entre les défaillances et les emprunteurs.

A. Óskarsdóttir et al. *Bra* surveillance fortement accru. Cet effet, dans le cas des prêts agricoles que nous étudions, est plus fort dans le réseau régional que dans le réseau de produits, ce qui montre également que les événements survenant dans une zone géographique sont plus importants que les événements affectant un produit unique. Dans ce contexte, une sécheresse est plus importante qu'une épidémie. Ces informations précédemment inconnues peuvent être très utiles pour concevoir des mesures préventives visant à atténuer les pertes en cas d'événements catastrophiques, et montrent l'importance d'une approche de réseau multicouche. En ce qui concerne les variables du PageRank multicouche, la plus complète (Bi- part_combined) est la variable la plus prédictive. Cela indique que le réseau multicouche permet une propagation plus riche du risque. Elle est suivie par la variable Aggregate, qui est un calcul de PageRank sur le réseau "agrégé", ce qui indique également que la connectivité du réseau est la source du pouvoir prédictif.

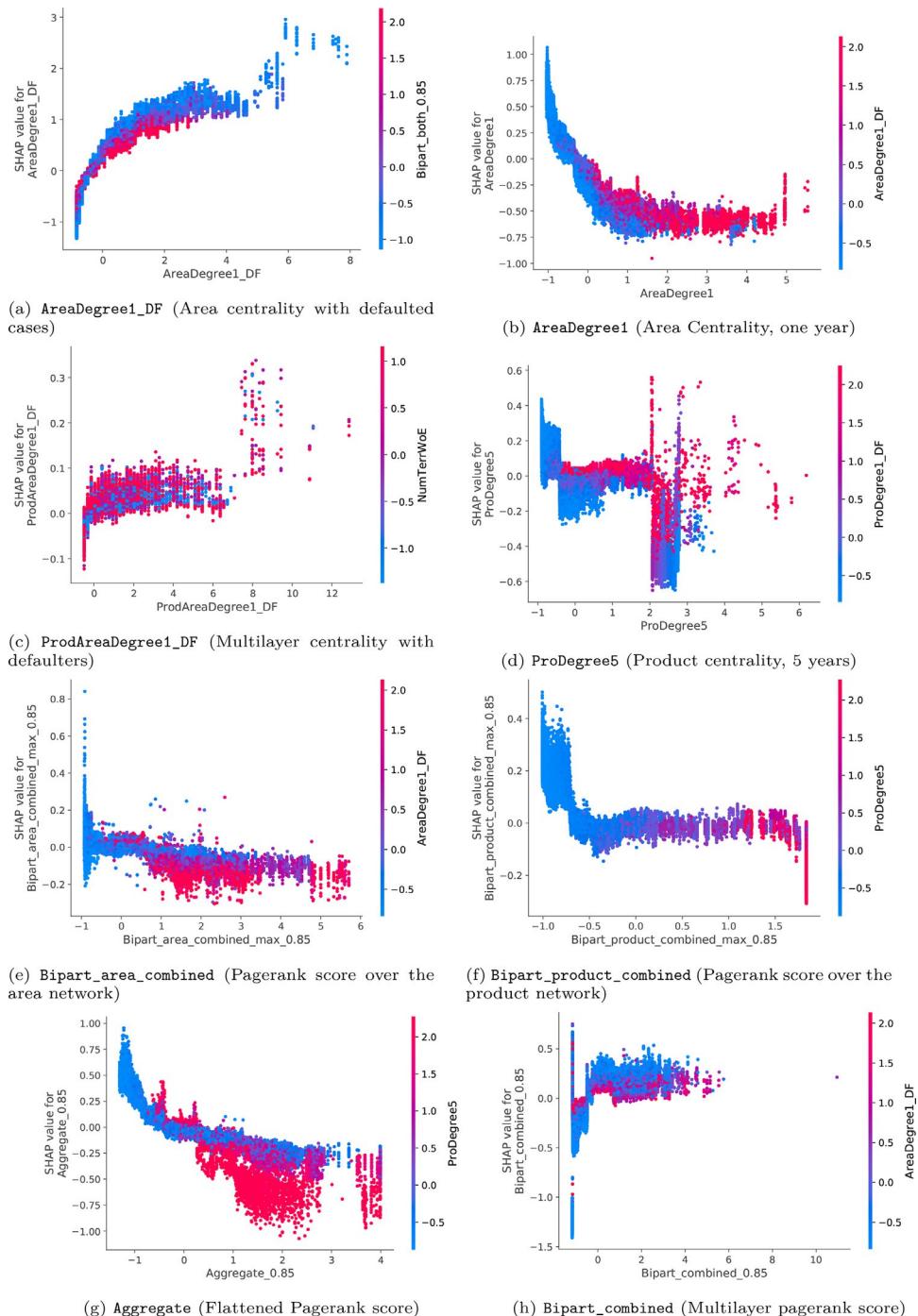


Fig. 5. Diagrammes de dépendance pour les variables du réseau. La couleur indique la valeur de la variable la plus proche par corrélation.

Nous avons établi que les variables de réseau sont significatives et prédictives. La section suivante étudie en profondeur le comportement des variables de réseau les plus significatives et les enseignements qui peuvent en être tirés ().

4.5. Interprétation du modèle

Dans cette section, nous analyserons comment les variables clés du réseau interagissent dans le modèle pour faire des prédictions. Pour ce faire, nous utiliserons les diagrammes de dépendance de TreeShap sur le modèle XGBoosting Baseline + All Network Variables. Les diagrammes de dépendance permettent d'étudier l'impact de chaque variable sur la prédiction (à l'aide de

la valeur de Shapley).

Les variables sont utilisées pour étudier comment elles influencent une prédiction vers une valeur de 0 ou 1), et également pour étudier comment chaque variable est liée aux valeurs du prédicteur le plus proche (en termes de corrélation). Cela nous permet de brosser une image très détaillée de la manière dont une prédiction est calculée.

Le premier résultat intéressant est que la plupart des effets de réseau sont nettement non linéaires. Les valeurs extrêmes des variables entraînent de fortes variations dans les prédictions. Cela signifie que les emprunteurs fortement (légèrement) connectés sont plus (moins) exposés au risque de défaillance, mais que les emprunteurs modérément connectés n'ont généralement pas un risque de réseau aussi élevé.

Si l'on se concentre sur des segments particuliers de variables, les première et deuxième lignes ([Fig. 5a-d](#)) montrent un comportement combiné très intéressant.

Les deux premières présentent des complémentarités, AreaDegree1_DF augmentant rapidement le risque de défaillance à mesure que la valeur de la variable augmente, ce qui est contrasté par un risque décroissant à mesure que la valeur de la variable opposée (AreaDegree1) augmente. Cela donne un résultat très intéressant : en général, les personnes très connectées ont un

meilleur comportement que l'emprunteur moyen, mais les personnes très connectées à des emprunteurs ayant déjà fait défaut courrent un risque plus élevé. Les figures 5c et d semblent suggérer que l'effet de la centralité globale est plus fort que la centralité en ce qui concerne les emprunteurs défaillants, de sorte qu'une communauté plus forte soutient davantage les individus qu'un ralentissement propagé ne les expose au risque. Les valeurs de Shapley suggèrent également qu'une connexion très élevée avec les défaillants est très prédictive de la défaillance.² Ces deux variables de centralité présentent également un comportement fortement non linéaire et, pour les valeurs extrêmes des variables, elles montrent également une forte capacité prédictive.

Les troisième et quatrième rangées de figures montrent l'impact des scores PageRank. Là encore, leur comportement n'est pas linéaire et les valeurs situées autour du centre de la distribution ne fournissent que peu d'informations, les extrêmes étant de meilleurs prédicteurs de défaillance ou de non-défaillance selon l'extrémité de la variable. Le score PageRank du réseau agrégé semble être un mauvais indicateur pour les valeurs moyennes, alors que tous les scores (et en particulier le score multicouche, figure 5h) semblent être d'excellents indicateurs d'un bon remboursement pour les valeurs faibles. Comme les scores PageRank calculent l'exposition au risque, les emprunteurs qui présentent la plus faible probabilité de défaillance sont ceux qui, après propagation, ne semblent pas être exposés.

La synthèse des deux analyses aboutit à une conclusion intéressante : La centralité montre que le degré de connexion d'un emprunteur avec les mauvais payeurs est un bon indicateur du risque de défaillance, tandis que les scores de propagation montrent que même si l'emprunteur est connecté, si sa proximité avec les mauvais payeurs est suffisamment grande (comme mesuré par le score PageRank), son risque reste faible. C'est en utilisant les deux ensembles de variables combinés que la nouvelle capacité de prédiction apparaît.

5. Conclusions

Dans cet article, nous avons présenté un cadre permettant de transformer un ensemble de données en un réseau multicouche avec des réseaux intra-couche bipartites, en utilisant deux variables de connexion ou plus dans l'ensemble de données. Ce cadre peut être appliqué à tout ensemble de données comportant des variables de connexion raisonnables afin d'obtenir un réseau multicouche interconnecté pour un ensemble de données apparemment dépourvu de structure de réseau. En outre, nous avons mis au point une mesure de centralité PageRank multicouche personnalisée qui peut être utilisée pour propager l'influence - ou un autre effet - d'un ensemble de nœuds sources à travers le réseau et ses couches, et ainsi étudier le degré d'exposition des nœuds à l'influence. Enfin, nous avons appliqué notre cadre et notre mesure de centralité à un ensemble de données de prêts agricoles, dans lequel nous avons connecté les emprunteurs à l'aide de variables décrivant les produits et les districts. Nous avons créé des modèles d'évaluation du crédit avec des variables du réseau en plus des variables traditionnelles et nous avons constaté une augmentation significative des performances.

Nous avons utilisé notre nouvelle mesure de centralité PageRank multicouche personnalisée pour calculer des scores qui quantifient le degré d'exposition des nœuds à l'influence des nœuds sources. Comme les scores sont calculés pour tous les nœuds du réseau, il est possible d'évaluer le degré de risque des observations et des variables

de connexion, en l'occurrence les produits et les districts [11]. En outre, sur la base du réseau et des scores PageRank, nous pouvons extraire diverses variables du réseau, telles que le nombre d'emprunteurs dans le même district et le score PageRank des nœuds voisins. Ces variables sont des propriétés des nœuds eux-mêmes, que nous avons ensuite ajoutées à l'ensemble de données sur les prêts et l'avons ainsi enrichi

² Il semble que, au moins empiriquement et pour ce problème particulier, le vieil adage "les amis valent plus que l'argent" soit valable... à moins que la plupart des gens autour de vous ne soient des mauvais payeurs.

Maria Óskarsdóttir et Cristián Bravo avec des informations provenant du réseau. Dans notre cas, les variables de connexion étaient catégoriques et comportaient des centaines de niveaux, ce qui les rendait difficiles à inclure directement dans le modèle d'évaluation du crédit. Grâce à la construction de notre réseau, leur effet a pu être intégré de manière transparente.

Une expérience approfondie montre qu'en général, les variables du réseau sont utiles pour prédire les défaillances, et que les approches multicouches sont plus utiles que les approches monocouches. Nous obtenons des améliorations de plus de 10 % sur les mesures AUC par rapport aux modèles de base. Il est intéressant de noter que les améliorations semblent être fortement non linéaires, ce qui suggère que des approches d'apprentissage automatique plus puissantes doivent être utilisées pour obtenir des capacités prédictives optimales à partir de ces informations. Enfin, et conformément à la littérature, les variables comportementales sont les prédicteurs de défaillance les plus puissants, mais des améliorations d'environ 2,5 % peuvent être obtenues en utilisant des variables de réseau.

L'utilisation de mesures d'interprétabilité sur les variables de réseau apporte également des informations intéressantes qui n'auraient pas été révélées si ces variables n'avaient pas été incluses. Les variables de réseau présentent des effets non linéaires et complémentaires qui doivent être soigneusement étudiés pour comprendre comment le risque de défaillance se propage dans les réseaux et expose les emprunteurs sains au risque. Ces informations peuvent être utilisées par les prêteurs et les régulateurs pour prendre des mesures préventives lorsqu'ils constatent une détérioration des portefeuilles : les emprunteurs à risque peuvent être contactés avec des offres de refinancement ou des programmes de soutien peuvent être mis en place dans des conditions économiques particulièrement difficiles.

Nous pensons que les résultats de ce travail démontrent clairement l'existence d'un risque de défaillance corrélé et la manière dont ce risque de défaillance peut être explicitement pris en compte en utilisant des réseaux bipartites multicouches. Ces mesures peuvent être utilisées dans des travaux futurs pour estimer des modèles prédictifs et économétriques détaillés sur les emprunteurs de détail dans de nombreux domaines. Notre travail propose d'autres orientations pour la recherche future. Tout d'abord, les méthodes d'apprentissage profond pour les données nettes, appelées réseaux neuronaux graphiques, ont récemment progressé. Il serait intéressant de développer une telle méthode pour les réseaux bipartites multicouches qui peuvent apprendre les scores de crédit - ou toute étiquette de nœud - directement à partir de la structure du réseau, et d'étudier la comparaison avec la méthode présentée dans le présent document. Deuxièmement, la mesure de centralité multicouche PageRank peut être étudiée plus en détail. Cela inclut l'effet de la valeur de redémarrage, de l'adhérence ainsi que de la représentation de la source d'influence sur le classement des nœuds. Dans le présent document, nous l'avons appliquée à un réseau à deux couches, mais la complexité augmente avec le nombre de couches, ce qui pourrait faire l'objet d'une étude. En outre, la mesure pourrait faire l'objet de tests de résistance en termes de risque de crédit. Enfin, nous aimerions appliquer ces méthodes à d'autres domaines que le risque de crédit, en particulier utiliser le cadre pour construire des réseaux multicouches avec divers ensembles de données et y mesurer la centralité du lien personnalisé. Cela contribuerait au domaine de la science des réseaux en l'étendant à des contextes où la structure du réseau n'est pas explicite et permettrait à d'autres de récolter les avantages de la science des réseaux.

Déclaration de contribution des auteurs du CRedit

María Óskarsdóttir : Conceptualisation, Méthodologie, Logiciel, Analyse formelle, Réécriture - version originale, Réécriture - relecture et édition, Visualisation. **Cristián Bravo** : Conceptualisation, Méthodologie, Logiciel, Analyse formelle, Réécriture - version originale, Réécriture - relecture et édition,

Remerciements

Le deuxième auteur remercie le [Conseil de recherches en sciences naturelles et en génie du Canada](#) (CRSNG) pour son soutien [subvention à la découverte RGPIN-2020-07114]. Cette recherche a été réalisée, en partie, grâce au financement du Programme des chaires de recherche du Canada.

Annexe A. Variables du réseau

Tableau A.3

Liste des variables extraites des réseaux.

Variable	Description de la variable
ProDegree1	Nombre d'emprunteurs ayant utilisé le même produit au cours de l'année écoulée
ProDegree1_DF	Nombre d'emprunteurs en défaut de paiement avec le même produit au cours de l'année écoulée
ProDegree5	Nombre d'emprunteurs ayant utilisé le même produit au cours des cinq dernières années
ProDegree5_DF	Nombre d'emprunteurs en défaut de paiement avec le même produit au cours des cinq dernières années
DistDegree1	Nombre d'emprunteurs dans le même district au cours de l'année écoulée
DistDegree1_DF	Nombre d'emprunteurs défaillants dans le même district au cours de l'année écoulée
DistDegree5	Nombre d'emprunteurs dans le même district au cours des cinq dernières années
DistDegree5_DF	Nombre d'emprunteurs défaillants dans le même district au cours des cinq dernières années
Degré de surface1	Nombre d'emprunteurs dans la même région au cours de l'année écoulée
AreaDegree1_DF	Nombre d'emprunteurs défaillants dans la même zone au cours de l'année écoulée
Degré de surface5	Nombre d'emprunteurs dans la même région au cours des cinq dernières années
AreaDegree5_DF	Nombre d'emprunteurs en défaut de paiement dans la même zone au cours des cinq dernières années
ProdDistDegree1	Nombre d'emprunteurs dans le même district et avec le même produit au cours de la dernière année
ProdDistDegree1_DF	Nombre d'emprunteurs en défaut de paiement dans le même district et avec le même produit au cours de la dernière année
la dernière année ProdDistDegree5	Nombre d'emprunteurs dans le même district et avec le même produit au cours des cinq dernières années
ProdDistDegree5_DF	Nombre d'emprunteurs en défaut de paiement dans le même district et avec le même produit au cours des cinq dernières années
des cinq dernières années ProdAreaDegree1	Nombre d'emprunteurs dans la même zone et avec le même produit au cours de la dernière année
ProdAreaDegree1_DF	Nombre d'emprunteurs en défaut de paiement dans la même zone et avec le même produit au cours de la dernière année
dernière année ProdAreaDegree5	Nombre d'emprunteurs dans la même zone et avec le même produit au cours des cinq dernières années
ProdAreaDegree5_DF	Nombre d'emprunteurs en défaut de paiement dans la même région et avec le même produit au cours des cinq dernières années
cinq dernières années Bipart_intra	Score PageRank personnalisé avec la matrice d'influence intra
Bipart_inter	Score PageRank personnalisé avec la matrice d'influence inter
Bipart_combiné	Score PageRank personnalisé avec la matrice d'influence combinée
Bipart_product_intra_max	Le score PageRank personnalisé du produit de l'emprunteur ayant le score le plus élevé, calculé avec la méthode intra matrice d'influence
Bipart_product_inter_max	Le score PageRank personnalisé du produit de l'emprunteur ayant le score le plus élevé, calculé avec la méthode inter matrice d'influence
Bipart_product_combined_max	Score PageRank personnalisé du produit de l'emprunteur ayant le score le plus élevé, calculé à l'aide des scores combinés de matrice d'influence
Bipart_district_intra_max	Le score PageRank personnalisé du district de l'emprunteur ayant le score le plus élevé, calculé avec la méthode intra matrice d'influence
Bipart_district_inter_max	Le score PageRank personnalisé du district de l'emprunteur ayant le score le plus élevé, calculé avec la méthode inter matrice d'influence
Bipart_district_combiné_max	Le score PageRank personnalisé du district de l'emprunteur ayant le score le plus élevé, calculé à l'aide des scores combinés de matrice d'influence
Bipart_area_intra_max	Le score PageRank personnalisé de la zone de l'emprunteur ayant le score le plus élevé, calculé avec l'influence intra matrice
Bipart_area_inter_max	Le score PageRank personnalisé de la zone de l'emprunteur ayant le score le plus élevé, calculé avec l'inter influence matrice
Bipart_area_combined_max	Score PageRank personnalisé de la zone de l'emprunteur ayant le score le plus élevé, calculé à l'aide des scores combinés de matrice d'influence

Annexe B. Description des variables hors réseau et statistiques descriptives

Tableaux B.4 et B.5.

Tableau B.4

Description des variables non basées sur le

réseau. Variable	Description de la
variable	
Garant	Le prêt est-il assorti d'un garant ?
LoanAmt	Montant du prêt (masqué)
TermMonths	Durée en mois
HectaresTerrain	Surface du terrain où a lieu la production principale
HRB	Rendement par ha.
Âge de l'emprunteur	Âge de l'emprunteur
Nombre de garanties	Nombre de titres attachés à l'opération ValCol Valeur des titres (masquée)
num_products	Nombre de produits vendus par l'emprunteur
Arrears15H	Pourcentage de paiements en retard de plus de 15 jours pour les prêts
précédents NegativeMoveH	Si un mouvement négatif s'est produit pour des prêts accordés précédemment
NumOtherLoans	Nombre de prêts accordés avant le prêt actuel
NumConcurrLoans	Nombre de prêts encore en cours de remboursement avant le prêt
actuel IfOtherLoans	L'emprunteur a-t-il déjà bénéficié d'un prêt ?
IfConcurrentLoans	L'emprunteur a-t-il un prêt en cours ? SiArrear15Current L'un des prêts en cours a-t-il fait l'objet d'un retard de paiement de plus de 15 jours ? MaximumArrearHistoric
en cours a-t-il fait l'objet d'un retard de paiement de plus de 15 jours ? MaximumArrearHistoric	Nombre maximum de jours d'arriérés pour tous les prêts accordés
précédemment Défaut de paiement	Défaut de paiement à 90 jours pendant la durée du prêt

Tableau B.5
Statistiques descriptives des variables hors réseau.

Variable	Moyenne	Std. Dev.	Min.	Max.
Garant	0.08	0.28	0	1
Montant du prêt	42.88	77.27	0	7903
TermeMois	30.96	11.43	13	71
HectaresTerrain	69.54	273.67	0	15411
HRB	4.22	5.10	0	74
L'âge	52.83	13.92	17	110
NombreCollatéral	0.50	0.53	0	4
ValCol	57.97	211.23	0	27689
nombre de produits	1.52	0.84	1	9
Arriérés15H	0.19	0.29	0	1
NegativeMoveH	0.48	0.50	0	1
NumOtherLoans	6.89	5.98	1	55
NumConcurrLoans	1.15	1.12	0	11
SiAutresPrêts	0.89	0.32	0	1
SiPrêtsconcurrents	0.19	0.40	0	1
SiArrear15Courant	0.15	0.36	0	1
MaximumArrearHistorique	100.38	237.93	0	3408
Défaut	0.12	0.33	0	1

Références

- [1] Amoroso N, La Rocca M, Bruno S, Maggipinto T, Monaco A, Bellotti R, et al. Réseaux multiplexes pour le diagnostic précoce de la maladie d'Alzheimer. *Front Ag-ing Neurosci* 2018;10:365.
- [2] Bagavathi A, Krishnan S. Multi-Net : a scalable multiplex network embedding framework. In : International conference on complex networks and their applications. Springer ; 2018. p. 119-31.
- [3] Bai C, Shi B, Liu F, Sarkis J. Banking credit worthiness : evaluating the complex relationships. *Omega* 2019;83:26-38.
- [4] Barabási A-L, et al. La science des réseaux. Cambridge university press ; 2016.
- [5] Comité de Bâle sur le contrôle bancaire. An explanatory note on the basel II IRB risk weight functions (Note explicative sur les fonctions de pondération des risques de Bâle II IRB). Tech. Rep. Banque des règlements internationaux ; 2005.
- [6] Behzadi G, O'Sullivan MJ, Olsen TL, Zhang A. Agribusiness supply chain risk management : a review of quantitative decision models. *Omega* 2018;79:21-42. doi:[10.1016/j.omega.2017.07.005](https://doi.org/10.1016/j.omega.2017.07.005).
- [7] Boldi P, Santini M, Vigna S. A deeper investigation of PageRank as a function of the damping factor. In : Dagstuhl seminar proceedings. Schloss Dagstuhl-Leibniz-Zentrum für Informatik ; 2007.
- [8] Bookstaber R, Kenett DY. Looking deeper, seeing more : a multilayer map of the financial system (Regarder plus loin, voir plus loin : une carte multicouche du système financier). *OFR Brief* 2016;16(6):1-12.
- [9] Botelho J, Antunes C. Combining social network analysis with semi-supervised clustering : a case study on fraud detection. In : Proceeding of mining data semantics (MDS'2011) in conjunction with SIGKDD. San Diego, CA, USA : Citeseer ; 2011. p. 1-7.
- [10] Bravo C, Maldonado S, Weber R. Octroi et gestion de prêts pour les micro-entrepreneurs : nouveaux développements et expériences pratiques. *Eur J Oper Res* 2013;227(2):358-66.
- [11] Bravo C, Óskarsdóttir M. Evolution of credit risk using a personalized PageRank algorithm for multilayer networks. In : Proceedings of third KDD workshop on machine learning in Finance, Joint with 26th ACM SIGKDD conference on knowledge discovery in databases. San Diego, CA, USA (en ligne) : ACM ; 2020. p. 8 pages.
- [12] Bressan M, Peserico E. Choose the damping, choose the ranking ? *J Discrete Algorithms* 2010;8(2):199-213.
- [13] Bui TD, Ravi S, Ramavajjala V. Neural graph learning : training neural networks using graphs. In : Proceedings of the eleventh ACM international conference on web search and data mining ; 2018. p. 64-71.
- [14] Cen Y, Zou X, Zhang J, Yang H, Zhou J, Tang J. Representation learning for attributed multiplex heterogeneous network. In : Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining ; 2019. p. 1358-68.
- [15] Cheng D, Niu Z, Zhang Y. Contagious chain risk rating for networked-guarantee loans. In : Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. San Diego, CA, USA (en ligne) : ACM ; 2020. p. 2715-23.
- [16] Cherian J, Sajeev G. An improved PageRank algorithm for multilayer networks. In : 2020 IEEE International conference on electronics, computing and communication technologies (CONECCT). IEEE ; 2020. p. 1-6.
- [17] De Domenico M, Porter MA, Arenas A. muxViz : a tool for multilayer analysis and visualization of networks. *J Complex Netw* 2015;3(2):159-76.
- [18] De Domenico M, Solé-Ribalta A, Cozzo E, Kivelä M, Moreno Y, Porter MA, et al. Mathematical formulation of multilayer networks. *Phys Rev X* 2013;3(4):041022.
- [19] De Domenico M, Sol-Ribalta A, Omodei E, Gmez S, Arenas A. Ranking in interconnected multilayer networks reveals versatile nodes. *Nat Commun* 2015;6:6868.
- [20] Feinstein Z. Obligations with physical delivery in a multilayered financial network. *SIAM J Financ Math* 2019;10(4):877-906. doi:[10.1137/18M1194729](https://doi.org/10.1137/18M1194729).
- [21] Fenech JP, Vosgha H, Shafik S. Loan default correlation using an Archimedean copula approach : a case for recalibration. *Econ Model* 2015;47:340-54. doi:[10.1016/j.econmod.2015.03.001](https://doi.org/10.1016/j.econmod.2015.03.001).
- [22] Garas A. Réseaux interconnectés. NY, USA : Springer ; 2016.
- [23] Getoor L. Link-based classification. In : Advanced methods for knowledge discovery from complex data. Springer ; 2005. p. 189-207.
- [24] Gomez S, Diaz-Guilera A, Gomez-Gardenes J, Perez-Vicente CJ, Moreno Y, Arenas A. Diffusion dynamics on multiplex networks. *Phys Rev Lett* 2013;110(2):028701.
- [25] Grover A, Leskovec J. node2vec : scalable feature learning for networks. In : Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining ; 2016. p. 855-64.
- [26] Gupta S, Kumar P. A constrained agglomerative clustering approach for unipartite and bipartite networks with application to credit networks. *Inf Sci* 2020.
- [27] Halu A, Mondragón RJ, Panzarasa P, Bianconi G. Multiplex PageRank. *PloS one* 2013;8(10).
- [28] Iacovacci J, Bianconi G. Extracting information from multiplex networks. *Chaos* 2016;26(6):065306.
- [29] Iacovacci J, Rahmede C, Arenas A, Bianconi G. Functional multiplex PageRank. *EPL* 2016;116(2):28004.
- [30] Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. *J Complex Netw* 2014;2(3):203-71.
- [31] Kwak H, Lee C, Park H, Moon S. Qu'est-ce que twitter, un réseau social ou un média d'information ? In : Proceedings of the 19th international conference on world wide web. Raleigh, NC, USA : ACM ; 2010. p. 591-600.
- [32] Langville AN, Meyer CD. Deeper inside PageRank. *Internet Math* 2004;1(3):335-80.
- [33] Lessmann S, Baesens B, Seow H-V, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring : an update of research. *Eur J Oper Res* 2015;247(1):124-36. doi:[10.1016/j.ejor.2015.05.030](https://doi.org/10.1016/j.ejor.2015.05.030).
- [34] Li J, Chen C, Tong H, Liu H. Multi-layered network embedding. In : Proceedings of the 2018 SIAM international conference on data mining. SIAM ; 2018. p. 684-92.
- [35] Lohmann G, Margulies DS, Horstmann A, Pleger B, Lepsién J, Goldhahn D, et al. Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PloS one* 2010;5(4):e10232.
- [36] Lundberg SM, Lee S-I. Une approche unifiée de l'interprétation des prédictions des modèles. In : Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, éditeurs. Advances in neural information processing systems 30. Curran Associates, Inc ; 2017. p. 4765-74.
- [37] Macskassy SA, Provost F. Classification in networked data : a toolkit and a univariate case study. *J Mach Learn Res* 2007;8(May):935-83.
- [38] Min W, Tang Z, Zhu M, Dai Y, Wei Y, Zhang R. Behavior language processing with graph based feature generation for fraud detection in online lending. In : Proceedings of workshop on misinformation and misbehavior mining on the web. Marina Del Rey, CA, USA : ACM ; 2018. p. 1-8.
- [39] Montagna M, Kok C. Multi-layered interbank model for assessing systemic risk. Document de travail 1944. Banque centrale européenne ; 2016.
- [40] Ning N, Wu B, Peng C. Representation learning based on influence of node for multiplex network. In : 2018 IEEE Third international conference on data science in cyberspace (DSC). IEEE ; 2018. p. 865-72.
- [41] Óskarsdóttir M, Bravo C, Sarraute C, Vanthienen J, Baesens B. The value of big data for credit scoring : enhancing financial inclusion using mobile phone data and social network analytics. *Appl Soft Comput* 2019;74:26-39.

- [42] Óskarsdóttir M, Bravo C, Verbeke W, Sarraute C, Baesens B, Vanthienen J. Social network analytics for churn prediction in telco : model building, evaluation and network architecture. *Expert Syst Appl* 2017;85:204-20.
- [43] Óskarsdóttir M, Cornette S, Deseure F, Baesens B. Inductive representation learning on feature rich complex networks for churn prediction in telco. In : International conference on complex networks and their applications. Springer ; 2019. p. 845-53.
- [44] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking : bringing order to the web. Tech. Rep. Stanford InfoLab ; 1999.
- [45] Peck Christen R, Pearce D, Rubio F, Acevedo JP, Brar A, Ayee G, et al. Managing risks and designing products for agricultural microfinance : features of an emerging model. Banque mondiale ; 2012.
- [46] Pedroche F, Romance M, Criado R. A biplex approach to PageRank centrality : from classic to multiplex networks. *Chaos* 2016;26(6):065301.
- [47] Poledna S, Molina-Borboa JL, Martínez-Jaramillo S, Van Der Leij M, Thurner S. La nature de réseau multicouche du risque systémique et ses implications pour les coûts des crises financières. *J Financ Stab* 2015;20:70-81.
- [48] Thomas L, Crook J, Edelman D. Credit scoring and its applications. 2e éd. USA : SIAM ; 2017.
- [49] Thomas LC, Oliver RW, Hand DJ. Une enquête sur les questions relatives à la recherche sur la modélisation du crédit à la consommation. *J Oper Res Soc* 2005;56(9):1006-15.
- [50] Thurner S, Poledna S. DebtRank-transparency : contrôle du risque systémique dans les réseaux financiers. *Sci Rep* 2013;3:1888.
- [51] Tu X, Jiang G-P, Song Y. Multiplex PageRank in multilayer networks considering shunt. In : International conference on science of cyber security. Springer ; 2019. p. 47-58.
- [52] Van Vlasselaer V, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B. GOTCHA ! Détection de fraude basée sur les réseaux pour la fraude à la sécurité sociale. *Manage Sci* 2017;63(9):3090-110.
- [53] Zhang D, Yin J, Zhu X, Zhang C. Apprentissage de la représentation des réseaux : une enquête. *IEEE Trans Big Data* 2018.
- [54] Zhang H, Qiu L, Yi L, Song Y. Scalable multiplex network embedding. In : IJCAI, vol. 18 ; 2018. p. 3082-8.
- [55] Zhong Y, Shu J, Xie W, Zhou Y-W. Politiques optimales de crédit commercial et de réapprovisionnement pour la conception de réseaux de chaînes d'approvisionnement. *Omega* 2018;81:26-37. doi:[10.1016/j.omega.2017.09.006](https://doi.org/10.1016/j.omega.2017.09.006).
- [56] Zhou J., Cui G., Zhang Z., Yang C., Liu Z., Wang L., et al. Graph neural networks : a review of methods and applications. *arXiv preprint arXiv:181208434* 2018.