RECTO : REcommandation diminuant la Congestion par Transport Optimal

Guillaume Bied[1] [2], Elia Pérennès[1], Solal Nathan [2], Victor Alfonso Naya [2], Philippe Caillou [2], Bruno Crépon [1], Christophe Gaillac [3], Michèle Sebag [2]

[1]Centre de Recherche en Économie et Statistiques (CREST), France
[2]Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Université Paris-Saclay, France
[3]Nuffield College, Oxford University

APIA, 2023

Contexte: le projet VADORE

- Partenariat de long terme (en cours depuis 2018) entre :
  - Le service public de l'emploi français, Pôle emploi
  - Des chercheurs en économie au CREST (ENSAE) et à Oxford
    - Bruno Crépon, Christophe Gaillac, Elia Perennes
  - Des chercheurs en informatique au LISN (Université Paris-Saclay)
    - Philippe Caillou, Michèle Sebag, Guillaume Bied, Solal Nathan
- Objectif : développer et évaluer (par des tests A/B) des systèmes de recommandation pour le marché du travail, en tirant parti des larges volumes de données disponibles

Contexte: le projet VADORE

- De nombreux problèmes :
  - Démarrage à froid
  - Passage à l'échelle
  - Équité
  - Congestion
- Contributions académiques :
  - Présentation de l'algorithme développé pour éviter le démarrage à froid et passer à l'échelle

    Bied et al., à paraître (IJCAI 2023)
  - Travaux en cours (soumis) sur l'équité
- Succession de *beta tests* en partenariat avec Pôle emploi :
  - Auprès de 100 000 demandeurs en mars 2022
  - Auprès de 230 000 demandeurs en juin 2023, pour valider l'approche et comprendre une éventuelle aversion aux algorithmes
  - Campagnes à venir prévues pour évaluer l'algorithme et aller vers son internalisation à Pôle emploi

RECTO : REcommandation diminuant la Congestion par Transport Optimal

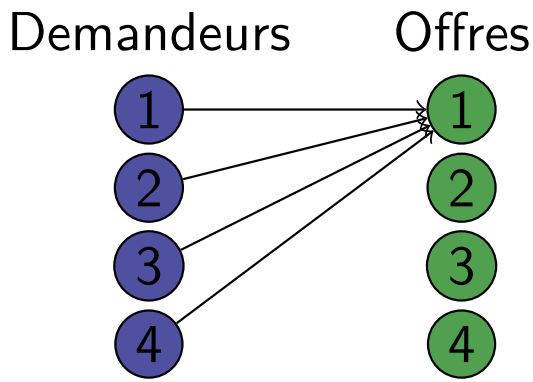
Guillaume Bied[1] [2], Elia Pérennès[1], Solal Nathan [2], Victor Alfonso Naya [2], Philippe Caillou [2], Bruno Crépon [1], Christophe Gaillac [3], Michèle Sebag [2]

[1]Centre de Recherche en Économie et Statistiques (CREST), France
[2]Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Université Paris-Saclay, France
[3]Nuffield College, Oxford University

APIA, 2023

Motivations

- *Cadre* : recommandation d'offres d'emploi à des demandeurs d'emploi
- *Un problème de recommandation particulier :*
  des biens **rivaux** $\rightarrow$ potentiel de **congestion**



- *Questions de recherche* :
  - La congestion constitue-elle un problème en pratique?
  - Comment peut-on la réduire?

Plan

État de l'art

- Recommandation & ressources humaines: challenges RecSys 16 & 17
  Volkovs et al., 2017
- Premiers travaux sur la congestion Gualdi et al., 2013
  - Lien avec le cadre de la recommandation réciproque (marché du travail, site de rencontres ...)
    Palomares et al., 2020
- Approches:
  - Optimisation sous contraintes Xia et al., 2019
    - à l'aide d'outils issus du transport optimal
      Chen et al., 2019; Li et al., 2019, Liu et al., 2019
  - Incorporer la popularité prédite des offres Borisyuk et al., 2017

Plan

Critères de performance

- Cadre: n demandeurs d'emploi, dont chacun reçoit une liste de k offres d'emploi
- Performance : recall@k
  - % d'utilisateurs du test t.q. leur futur contrat figure parmi les k premières recommandations
- Congestion : couverture@k
  - % d'offres recommandées au moins une fois (parmi les $n \times k$ recommandations émises)
- Un bon algorithme devrait être performant (haut recall) et engendrer peu de congestion (haute couverture)

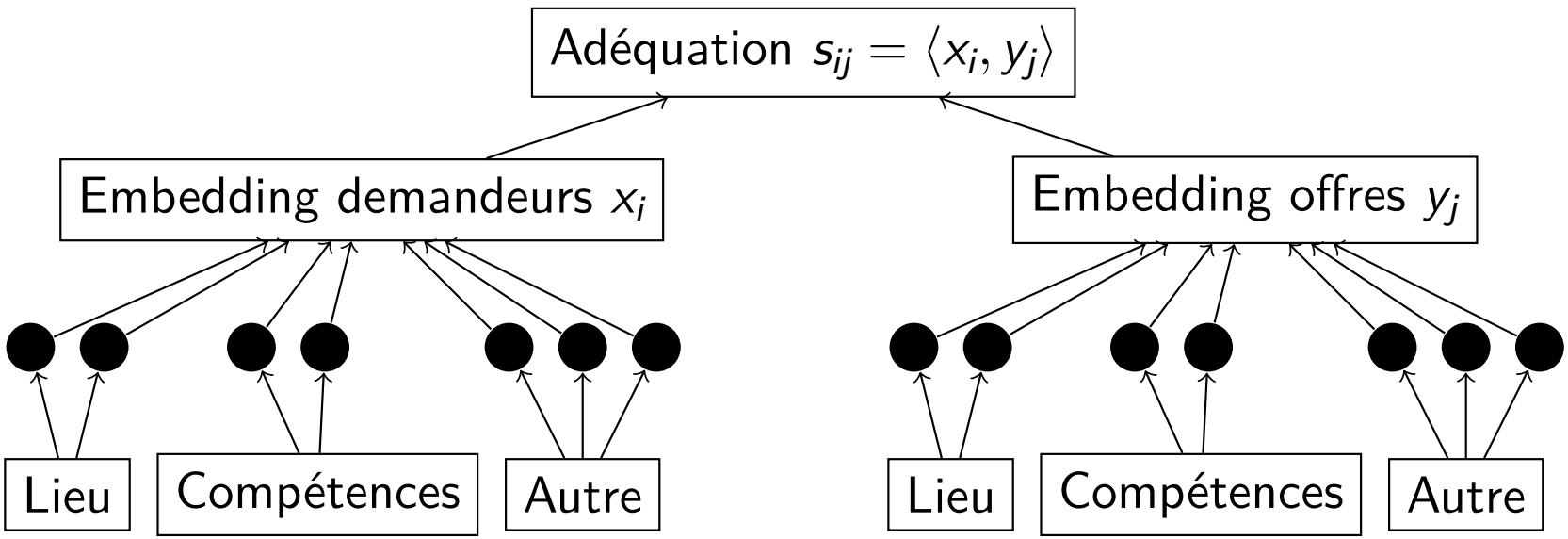
Données

- Labels : embauches
- Périmètre temporel : entraînement Feb.-Oct. 18; Test Nov. 18
- Région : Île-de-France
- Train : 1650k demandeurs d'emploi, 477k offres d'emploi, 43k embauches
- Test : 110k demandeurs d'emploi, 14k offres et 0.45k embauches
  - NB : 8 fois plus de demandeurs que d'offres !
- Représentation des demandeurs (resp. offres) dans $\mathbb{R}^{448}$ (resp. $\mathbb{R}^{582}$).
  - Demandeurs d'emploi : communes, métier, expérience, éducation, compétences, mobilité acceptée, critères de recherche
  - Offres : commune, métier, salaire, contrat, temps de travail, descriptif de l'offre et l'entreprise ...

Systèmes de recommandation étudiés

- XGBoost (XGB) : prédire si une paire est une embauche ou non à l'aide d'arbre boostés Volkovs et al., 2017
- Réseaux de neurones (NN), décrivant l'adéquation entre offres et demandeurs comme un produit scalaire entre embeddings, appris end-to-end avec une triplet loss:

$$\mathcal{L}(i, j, j') = [s_{ij} - s_{ij'} + 1]_+$$



Recommandation standard: premières leçons

Algorithme Recommandation	Recall (%) @1	Recall (%) @10	Couverture (%) @1	Couverture (%) @10
Aléatoire	0	0.21	99.95	100
XGB	9.62	**31.40**	12.94	**25.16**
NN	5.68	28.66	6.02	17.78

- La congestion constitue un réel problème
  - Exemple: environ 75% des offres n'apparaissent dans le top 10 d'aucun demandeur d'emploi

Plan

Transport optimal (TO)

- Entrées du problème :
  - Distribution uniforme discrète sur n utilisateurs μ
  - Distribution uniforme discrète sur m offres ν
  - Coût C_{ij} d'associer i et j
- Objectif : trouver un appariement de μ et ν au coût le plus faible :

$$\min_{\gamma \in M_{n,m}(\mathbb{R})} \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{i,j} C_{i,j} \tag{1}$$

$$\text{s.t.} \quad \gamma_{ij} \geq 0, \ \forall i,j; \quad \sum_{i} \gamma_{ij} = 1/n, \ \forall j; \quad \sum_{j} \gamma_{ij} = 1/m, \ \forall i$$

ID	1	2	3
1	5	10	40
2	10	12	15
3	5	20	100
4	10	30	20

Coûts

ID	1	2	3
1	0	0.25	0
2	0	0.08	0.17
3	0.25	0	0
4	0.08	0	0.17

Plan de TO γ

Illustration: $n = 4$ demandeurs (lignes), $m = 3$ offres (colonnes)

Régularisation entropique

- Cuturi (2013): relaxation à l'aide d'un terme entropique:

$$\min_{\gamma \in M_{n,m}(\mathbb{R})} \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{i,j} \left(C_{i,j} + \varepsilon \log(\gamma_{i,j}) \right) \tag{2}$$

avec ε pondérant la régularisation, toujours sous les contraintes:

$$\gamma_{ij} \geq 0, \ \forall i,j; \quad \sum_i \gamma_{ij} = 1/n, \ \forall j; \quad \sum_j \gamma_{ij} = 1/m, \ \forall i$$

ID	1	2	3
1	0.017	0.233	0
2	0	0.1	0.15
3	0.25	0	0
4	0.066	0	0.184

$\varepsilon = 1$

ID	1	2	3
1	0.082	0.087	0.081
2	0.073	0.079	0.098
3	0.1	0.096	0.054
4	0.079	0.071	0.1

$\varepsilon = 100$

Effet de la régularisation

L'algorithme RECTO

- *Question* : définir un coût C_{ij} de recommander i à j
- Étape 1: apprendre un score s_{ij} permettant de classer les offres j pour un demandeur i
- Étape 2:
  - Transformer s_{ij} en coût C_{ij}. Deux options considérées par la suite :
    - Linéaire : $C_{ij} \propto \beta s_{ij}$, seuillé pour les rangs > 1000
    - NDCG : C_{ij} calculé à partir de $NDCG(i, j)$
  - Résoudre $\gamma = OT(C_{ij}, \epsilon)$
  - Pour un demandeur d'emploi i, trier les offres par γ_{ij} décroissants

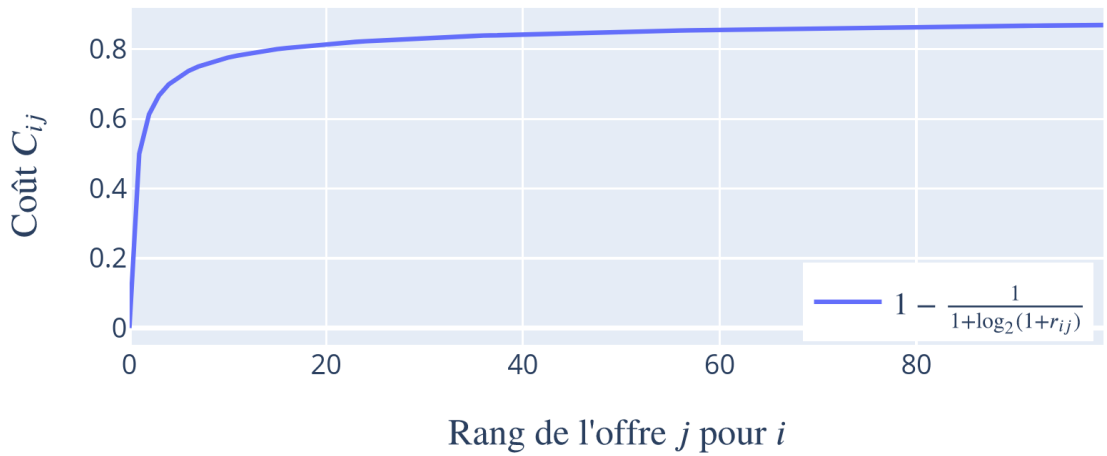


Figure: C_{ij} calculé à partir de NDCG

Validation empirique

- Les objectifs sont d'évaluer :
  - L'efficacité de RECTO en termes de compromis recall-congestion
  - Le rôle des hyper-paramètres
    - Définition du coût de transport C_{ij}
    - Poids de régularisation entropique ε
- NB : l'article considère aussi des données publiques sur des mariages pour une comparaison avec l'état de l'art (Li et al. 2018)

Résultats

Algorithme Recommandation		Recall (%) @1	Recall (%) @10	Couverture (%) @1	Couverture (%) @10
Aléatoire		0	0.21	99.95	100
XGB		9.62	31.40	12.94	25.16
$C_{i,j}$	ε				
Linéaire	1.0	4.81	21.99	21.61	31.76
Linéaire	0.1	2.18	15.31	27.54	41.24
Linéaire	0.01	**4.37**	**20.45**	**46.75**	**57.61**
NDCG	1.0	9.62	31.61	12.96	26.14
NDCG	0.1	8.97	25.38	14.69	30.84
NDCG	0.01	5.03	14.00	36.81	57.52

- On constate l'existence d'un trade-off entre couverture et recall
- De manière surprenante, baisser ε augmente la couverture (et réduit le recall)

Conclusion et perspectives

- Conclusion :
  - La congestion constitue un problème pratique
  - RECTO permet de réduire la congestion, au détriment du recall
- Perspectives :
  - Approche end-to-end de recommandations évitant la congestion
  - Beta tests en partenariat avec Pôle emploi

Conclusion et perspectives

- Conclusion :
  - La congestion constitue un problème pratique
  - RECTO permet de réduire la congestion, au détriment du recall
- Perspectives :
  - Approche end-to-end de recommandations évitant la congestion
  - Beta tests en partenariat avec Pôle emploi

Merci pour votre attention !

Contact : bied@lri.fr

Comparaison à l'état de l'art: Li et al., 2019

- Inspiration : TO vu comme un modèle de marchés décentralisés
  Shapley & Shubik, 1971; Galichon, 2016
- Sous cette hypothèse, Li et al. i/ apprennent une matrice de coût expliquant les données; ii/ utilisent les coûts appris pour matcher de nouvelles populations
- CAROT définit directement les coûts de modèles de recommandation standard
  PRO: i/ éviter de fortes hypothèses de modélisation du processus générateur des données; ii/ permet de s'appuyer sur l'état de l'art en recommandation
  CON: ii/ ajuster ϕ, avec $C_{i,j} = g(s_{i,j})$

Comparaison à l'état de l'art: Chen et al., 2019

(Dans le contexte de sites de rencontre)

- Pour des raisons de passage à l'échelle, Chen et al. procèdent de manière décentralisée; résolvent un problème de TO local dans le voisinage de chaque personne
- CAROT :
  - ne modélise pas les célibataires / chômeurs persistants comme des "sinks".
  - s'attaque au problème dans sa globalité (peut-être plus adapté pour éviter la congestion)

Jeu de données public : MAR (mariages)

- 2 475 hommes et 2 475 femmes, répartis en 50 groupes, décrits par 11 caractéristiques principalement ordinales.
- L'appariement 1 à 1 est décrit au niveau individuel
- Une matrice collaborative $M_{c,c'}$ indique la fraction d'appariements entre les hommes du groupe c et les femmes du groupe c'.
- Les résultats de référence sur MAR sont ceux de RIOT (Li et al., 2018), utilisant une factorisation basée sur SVD et itemKNN
- Indicateurs de performance niveau groupes: RMSE, MAE entre la matrice collaborative M au niveau des groupes et la matrice de recommandation estimée
- RECTO est également évalué au niveau individuel, à l'aide des indicateurs de performance usuels (recall, couverture)

Résultats MAR - niveau groupe

	PMF	SVD	itemKNN	RIOT	γ^{NN}	γ^{XGB}
RMSE	446.6± 9.86	441.4± 11.2	9.36± 0.12	9.12± 0.12	8.98± 0.17	8.89± 0.11
MAE	251.3± 6.00	249.2± 5.71	6.30± 0.03	5.98± 0.10	5.80± 0.13	5.79± 0.12

- Au niveau groupe, RECTO est compétitif avec RIOT et les baselines reportées par Li et al. (2018)

Résultats MAR - niveau individuel

Table: Résultats sur MAR au niveau individuel

	Algorithm	Recall (%) @1	Recall (%) @10	Couverture (%) @1	Couverture (%) @10
	ϕ Random	0.16	2.27	63.32	100
RECTO-XGB	ϕ XGB	7.93	27.88	48.55	98.69
RECTO-XGB	$\gamma^{XGB}, g = Id+, \varepsilon = 1.0$	8.05	28.41	49.77	99.18
RECTO-XGB	$\gamma^{XGB}, g = Id+, \varepsilon = 0.1$	8.01	27.02	72.73	100
RECTO-XGB	$\gamma^{XGB}, g = Id+, \varepsilon = 0.01$	6.47	23.77	96.05	100
RECTO-XGB	$\gamma^{XGB}, g = ndcg, \varepsilon = 1.0$	7.93	28.2	48.55	99.02
RECTO-XGB	$\gamma^{XGB}, g = ndcg, \varepsilon = 0.1$	8.10	25.72	59.42	100
RECTO-XGB	$\gamma^{XGB}, g = ndcg, \varepsilon = 0.01$	6.06	19.49	94.26	100
RECTO-NN	ϕ NN	3.82	15.50	46.27	98.00
RECTO-NN	$\gamma^{NN}, g = Id+, \varepsilon = 1.0$	2.84	14.32	38.86	92.47
RECTO-NN	$\gamma^{NN}, g = Id+, \varepsilon = 0.1$	3.94	15.46	70.12	100
RECTO-NN	$\gamma^{NN}, g = Id+, \varepsilon = 0.01$	3.78	15.46	93.48	100
RECTO-NN	$\gamma^{NN}, g = ndcg, \varepsilon = 1.0$	3.82	15.63	46.27	98.73
RECTO-NN	$\gamma^{NN}, g = ndcg, \varepsilon = 0.1$	4.23	13.87	57.99	99.91
RECTO-NN	$\gamma^{NN}, g = ndcg, \varepsilon = 0.01$	2.89	11.60	93.44	100

- XGB domine NN, tant en recall qu'en couverture
- En top-1, pour XGB, ε et la définition de C_{ij} permettent de naviguer de manière favorable le trade-off recall-couverture.
- Le test est une matrice de permutation: un "free lunch" recall-congestion est même possible sans contradiction