Fonctionnement et évolution de communautés en ligne à partir des rôles et comportements des contributeurs. Le cas de R et Pandas sur la plateforme numérique StackOverflow

Sébastien Delarre, Fabien Eloire, Maxime Morge, Antoine Nongaillard

Fonctionnement et évolution de communautés en ligne à partir des rôles et comportements des contributeurs. Le cas de R et Pandas sur la plateforme numérique *StackOverflow*

working paper, ne pas citer

Sébastien Delarre, Fabien Eloire, Maxime Morge, Antoine Nongaillard

Résumé :

Dans cet article nous analysons les échanges sur la plateforme StackOverlfow à l'aide de l'archive disponible en libre accès. Nous nous intéressons aux communautés R et Pandas, en focalisant sur les transferts d'utilisateurs d'un logiciel vers l'autre (Pandas étant de facture plus récente). Nos résultats montrent la rareté de ces transferts et le caractère très dual du support dans les groupes d'utilisateurs, dans lesquels des utilisateurs intéressés, souvent liés à des entreprises commerciales, ne font que répondre aux problèmes posés, tandis qu'une masse importante d'usagers ne fait qu'interroger à la cantonade, et reste dans des rôles passifs. Nous insistons sur le fait que ces réseaux numériques sont, en définitive, loin de se présenter comme des systèmes d'échanges généralisés dominés par la réciprocité et l'absence de hiérarchie.

Introduction

Cet article analyse la structuration des communautés et la migration des utilisateurs entre deux outils communément utilisés dans le monde des statistiques, et offrant des fonctionnalités proches : R et Pandas. Le logiciel R connait un plébiscite important dans le domaine universitaire depuis sa création en 2000. D'origine académique, il est un langage libre, permettant le développement et la diffusion de fonctions utilisateurs, chose impossible avec certains de ses concurrents, dont SAS, qui constituent des environnements fermés et quelques fois onéreux. La bibliothèque Pandas[1] est un équivalent de R, utilisable dans l'environnement de programmation Python, plus généraliste. Elle est apparue dix ans plus tard, en 2009, et présente de nombreuses similarités avec R. Dans les deux cas, il s'agit de logiciels massivement utilités pour le traitement de données au format *dataframe* (individus x variables), largement connus en SHS.

Pandas étant de facture plus récente que R, la situation des deux logiciels présente un cas intéressant de mise en concurrence de deux outils dans une communauté structurée d'usagers déjà constituée. Notre approche vise à questionner la structuration et les mécanismes de la conversion de certains utilisateurs, passant d'un logiciel à un autre, avec l'idée d'élargir notre propos aux implications épistémiques de ces conversions non ordonnées de communautés, aux moins en termes de

[1] https://pandas.pydata.org/about/

méthodes. En épistémologie le sujet a été discuté notamment par Feyerabend (1988), lequel insiste sur le rôle de la prolifération (des théories, méthodes et concepts) dans le développement des savoirs, critiquant toute forme de domination ou de monopole comme néfaste pour leur développement. Pour lui l'attachement aux idées nouvelles ne peut être provoqué que par des moyens irrationnels. Dans notre cas (R / Pandas), comment décrire le passage d'un groupe d'usagers d'un logiciel vers un autre ? Quels sont les mécanismes à l'œuvre ? En informatique, une plateforme de développement telle que *Github* fait l'objet de questionnements de ce type, insistant sur le rôle de la spéculation, hasardeuse, non dirigée, allant à l'encontre des dogmes imposés (le « *taboo of the fork* » Fuller et al. 2016). R et Pandas étant des logiciels centraux dans l'activité des chercheurs et des statisticiens en *data*, il nous a semblé intéressant de les prendre pour objet d'étude.

Le passage d'un logiciel (R) vers un autre plus récent (Pandas) n'est, certes, pas un phénomène comparable à une révolution de paradigme scientifique. Cependant l'utilisation de données d'archives massives issues du site web StackOverflow, à la fois exhaustives sur la période 2008-2020, et étendues à l'espace géographique international, nous semble être néanmoins l'occasion d'aborder le phénomène de conversion.

StackOverFlow (SO, cf. encadré 1) est une plateforme de type question - réponse (*Q&A Site*) largement utilisée dans le domaine de l'informatique. Créée en 2008 elle permet à ses utilisateurs de poser des questions ayant trait à la programmation, et de recevoir des réponses des membres plus expérimentés. Ces réponses font ensuite l'objet de votes en fonction de leur pertinence, et l'échange est ensuite archivé, laissé à disposition du grand public. SO est donc une forme de « super FAQ » (*Foire Aux Questions*) avec la particularité de viser à l'exhaustivité. Son intérêt pour nous est qu'elle constitue un nouveau modèle d'accès aux savoirs. SO concurrence donc l'édition scientifique d'ouvrages ou manuels d'auteurs, et l'enseignement. De par son fonctionnement, elle peut ainsi être lourde de conséquences sur le développement des connaissances pratiques et fondamentales à l'échelle mondiale. D'où notre intérêt à l'étudier.

L'exemple R / Pandas présente alors une opportunité rare de mesurer quantitativement un cas de « conversion épistémique » parce que SO met à disposition une vaste archive de données utilisateurs, le *Stack Exchange Data Dump* (cf. encadré 1). Issue directement de SO cette copie brute (*dump*) très détaillée comprend l'intégralité des échanges entre utilisateurs sur la plateforme. Disponible depuis 2008, nous l'utilisons sur une décennie complète et tentons d'observer l'effet de l'arrivée sur le marché d'un savoir concurrent (Pandas) sur un autre déjà installé (R). L'intérêt de cette source est également de mettre à l'épreuve certaines de nos habitudes d'analyse en tant que sociologues : en l'absence de variables causales d'attributs socio-démographiques, qu'est-ce que le sociologue a à dire sur de telles données numériques massives issues du web ?

Cadre théorique et hypothèses de recherche

Sur le plan de la théorie sociologique l'étude de ces données permet de questionner trois principes des modèles de *Threshold* développés en analyse de réseaux, qui semblent *a priori* tout à fait adaptés pour l'étude des communautés en ligne, dont celles générées par *SO*. Initialement (Schelling 1971, Granovetter 1978, Macy et Evtushenko 2020[2]) ces modèles cherchent à rendre compte des effets de seuils (*Critical Mass, Cascade Dynamics, Bandwagon Effect*) à partir desquels une

[2] Lesquels donnent une courte revue en préambule de la mise à jour stochastique du modèle initial

communauté bascule dans une pratique ou une opinion. Il s'agit de modèles d'action collective. Partant d'un groupe d'individus connectés entre eux, ces approches reposent sur une notion de niveaux de résistance individuels variables, dont la distribution dans le groupe provoque l'action collective en cascade, ou la bloque, si la cascade ne se produit pas. Elles visent à tenir compte de la structure en réseau du groupe : si les individus néophiles (ceux adoptant la nouveauté) forment une poche isolée dans le graphe d'ensemble, faiblement connectée aux autres, un changement global est peu probable (indépendamment de la taille de cette poche, et donc de la distribution marginale des seuils de résistance dans l'ensemble). Si ces néophiles ont des liens plus souvent hétérophiles, et que les seuils de résistance aux changements ne sont pas spécialement élevés dans le reste de la communauté, le changement peut advenir... En conséquence, la distribution statistique simple des seuils de tolérance dans le groupe n'a quasi aucune capacité prédictive, tant qu'on ne tient pas compte de la structure du réseau, c'est là le principal argument de ces techniques (« *cascades can be trapped in a local equilibrium that need not correspond to aggregate individual preferences* », Macy et Evtushenko 2020:632).

Nos données vont être l'occasion d'une triple critique de ces modèles. Premièrement, nous soulignons que, dans leur versions initiale et postérieure, ces approches ont eu tendance à partir d'une forme de vide social. Dans l'exemple donné par Granovetter notamment, l'acteur participe ou non à une action collective, mais le choix qu'il effectue est de l'ordre du « *oui »* ou « *je m'abstiens* » (d'entrer dans l'action collective). Nous pensons plus générale l'idée que l'acteur est face des choix positifs (action A v*s* action B). Une opinion, une pratique, une action se ne construit pas dans un néant, mais elle doit d'abord s'opposer à l'existant, ce qui semble assez négligé dans nombre d'approches. Avant de faire un modèle de la naissance d'un groupe social autour d'un nouveau *process*, il nous faut donc expliquer la scission initiale (c'est-à-dire le rôle et la place des instigateurs comme nous le verrons), autrement que par l'argument de la génération spontanée[3], ou par la manipulation expérimentale de quelques sommets dans un modèle à 100% issu de simulations. Qu'est-ce qui incite un acteur à changer d'opinion ou d'attitude face à la masse des autres, si tant est qu'il le fasse ? Nous verrons que des acteurs puissants aux propriétés réputationnelles très spécifiques, qui dépendent fortement de dynamiques extérieures au réseau étudié, jouent un rôle crucial en début de dynamique. Ils sont les précurseurs et les instigateurs du changement.

Deuxièmement nous rappelons que les *Threshold Models* en analyse de réseaux étudient le changement à l'intérieur d'un groupe social constitué, aux frontières fixes. Dans le cas qui nous occupe (une archive s'étalant sur 10 ans), nous constatons que les flux entrants d'acteurs sont permanents, et que cette dynamique mouvante a un rôle central - presque exclusif - sur la production collective de savoirs, et donc sur le changement. L'image que nous avons en tête est davantage celle d'une *course de relais intergénérationnelle* en continu. Elle est assez éloignée du postulat méthodologique d'une matrice totale et fixe d'individus en situation d'interconnaissance qui produisent du contenu, apprennent mutuellement et opèrent ensemble une conversion intra-biographique. Si cette approche de niveau acteur est fonctionnelle dans nombre de situations, nous voyons qu'elle ne s'adapte pas bien au cas empirique des communautés en ligne qui nous intéresse ici. Dès 1978 cependant, Granovetter soulignait d'ailleurs que l'entrée et/ou la sortie de nouveaux membres pouvait altérer fondamentalement le résultat de la simulation, sans pour autant mettre ce phénomène au cœur du débat. Le rôle joué par l'entrée et la sortie d'individus dans un groupe, n'est-

[3] « When cascades can be expected to fail because of seemingly "start-up problem", introducing a very small amount of noise at the individual level can lead to unexpected "spontaneous instigation" that may explain revolutionary surprises (…) " (Macy 2020:632).

il pas plus important sur le changement de normes, que les conflits entre individus déjà implantés autour de celles-ci ?

Troisièmement nous constatons que le perfectionnement des modèles en SNA (*Social Network Analysis*) a généralement visé à tenir compte de la structure du réseau, des cliques plus ou moins denses et des ponts existants entre ses différentes composantes (Chiang 2007 ; Galeotti et Goyal 2009). Or *StackOverflow* a la particularité - et l'ambition - de délivrer une information « pure et parfaite ». Quand un utilisateur questionne le moteur de recherche, ce dernier lui délivre une réponse basée sur l'ensemble de la base de données SO, sans tenir compte de ses liens d'affinités. L'idée centrale en SNA - que l'information est prise dans des poches de contacts et limite les perspectives et l'action de l'individu, et donc sa rationalité - se trouve, là encore, inadaptée, et certains de ses modèles ne sont pas adéquats[4]. La force de l'analyse de réseaux est de tenir compte de la structure des graphes, mais, dans certains cas, ce principe de la recherche n'emmène-t-elle pas l'analyse vers un schéma pas toujours approprié ?

Ainsi cette étude s'inscrit dans la perspective des *Threshold Models* en SNA, tout en étant à la recherche de nouveaux principes de généralité. Notre approche empirique se pose à l'écart de la tradition des simulations pures, et se propose de mieux tenir compte des particularités générées par les nouveaux dispositifs d'échanges collaboratifs, tels que *StackOverflow*. Nous verrons alors que le Web ne cadre pas nécessairement avec l'approche en termes de réseaux que nous avons tendance à lui imposer spontanément. Ayant en toile de fond ces questionnements théoriques, cet article est aussi l'occasion de proposer diverses méthodes expérimentales en instrumentant des données à la fois complexes, massives et pauvres en variables explicatives sociodémographiques traditionnelles, et en essayant de voir leur potentiel pour les SHS de façon plus générale.

Préparation des données et description du problème

L'objectif initial est la création d'une table de données unique dans laquelle les échanges sont identifiés à la façon de fils de discussion. De prime abord, le *Stack Exchange Data Dump* contient en effet des séries de contributions (questions et réponses) portant un identifiant commun, pour environ 1 268 600 observations. Le tableau 1 donne les effectifs bruts de l'archive après extraction, mais avant tout traitements. On constate une nette domination de l'étiquette R sur celle de Pandas et quelques scories très minoritaires (double étiquetage ou étiquetage hors sujet). La première version majeure de R date de 2000, Pandas fait son entrée en 2008, il est donc normal constater l'importance plus forte de R. L'objet de cet article est donc de s'interroger sur les modalités du basculement progressif de R vers Pandas, s'opérant dans la décennie 2010.

Tableau 1 : répartition initiale des étiquettes dans l'extraction

Tag de l'échange

[4] Ce qui n'est pas nécessairement le cas général en sociologie des réseaux puisque Macy 1991 par exemple teste différentes hypothèses sur la façon dont l'acteur observe le groupe lui-même (*groupwise serial interactions*), ou son réseau personnel spécifiquement.

Contenant l'étiquette R	846 918	66,8%
Contenant l'étiquette Pandas	420 192	33,1%
Contenant les deux étiquettes	1 466	0,1%
Ne contenant aucune des deux	24	0,0%
Total	**1 268 600**	**100,0%**

Source : *Stack Exchange Data Dump*

L'enjeu de cette section est d'observer le développement de ces deux outils d'analyse de données statistiques dans la communauté des utilisateurs de QO. On cherche donc d'abord à quantifier la porosité des deux communautés, et éventuellement les mécanismes de basculement d'un groupe épistémique ancien (R) vers un nouveau (Pandas).

Les contributeurs sont au nombre de 225 747 (pour 1,2 millions de publications environ). La répartition des contributions par individu montre une forte inégalité, avec une masse d'individus émettant peu (55% d'entre eux n'ont qu'une seule contribution – question ou réponse), et une vingtaine d'usagers (0.01% de l'ensemble) responsable d'un dixième de l'activité sur les deux domaines (R et Pandas). Le plus important d'entre eux émet environ 26 000 entrées sur une période de 6 ans (soit une dizaine de messages par jour, avec des amplitudes horaires pouvant correspondre à un cycle de travail normal). Ces individus prolixes seront donc centraux dans les analyses postérieures, afin de déterminer leur possible rôle moteur dans les conversions d'utilisateurs d'une communauté à l'autre. Ils auront une influence sur la méthodologie de création des séquences présentée dans la suite.

Encadré 1

Stack Overflow

Stack Overflow est une plateforme en ligne qui joue un rôle central dans la communauté mondiale des développeurs et des professionnels de l'informatique. Lancé en 2008 le site de questions-réponses SO est prééminent pour les codeurs de tous types, et concurrence - voire a annihilé - le marché de l'édition de manuels en langages de programmation. SO est une propriété du groupe Prosus, *actionnaire majoritaire du groupe chinois* Tencent, Prosus *est une filiale de la multinationale sud-africaine* Naspers*. Le principe fondamental de la plateforme est de permettre aux utilisateurs de poser des questions liées à la programmation, au développement logiciel, à la conception de sites web, à la gestion de bases de données et à d'autres sujets connexes. Les membres de la communauté, composée de développeurs expérimentés ou non, peuvent répondre à ces questions, offrant ainsi une solution aux difficultés des autres. La particularité de Stack Overflow réside dans son système de vote. Les réponses et les questions peuvent être évaluées par les utilisateurs en fonction de leur pertinence et de leur qualité. Les meilleures réponses, celles qui reçoivent le plus de votes, apparaissent ainsi en premier, offrant ainsi une hiérarchie claire de solutions aux problèmes courants. Les utilisateurs gagnent également des points de réputation en fonction de la qualité de leurs contributions.*

https://stackoverflow.com/

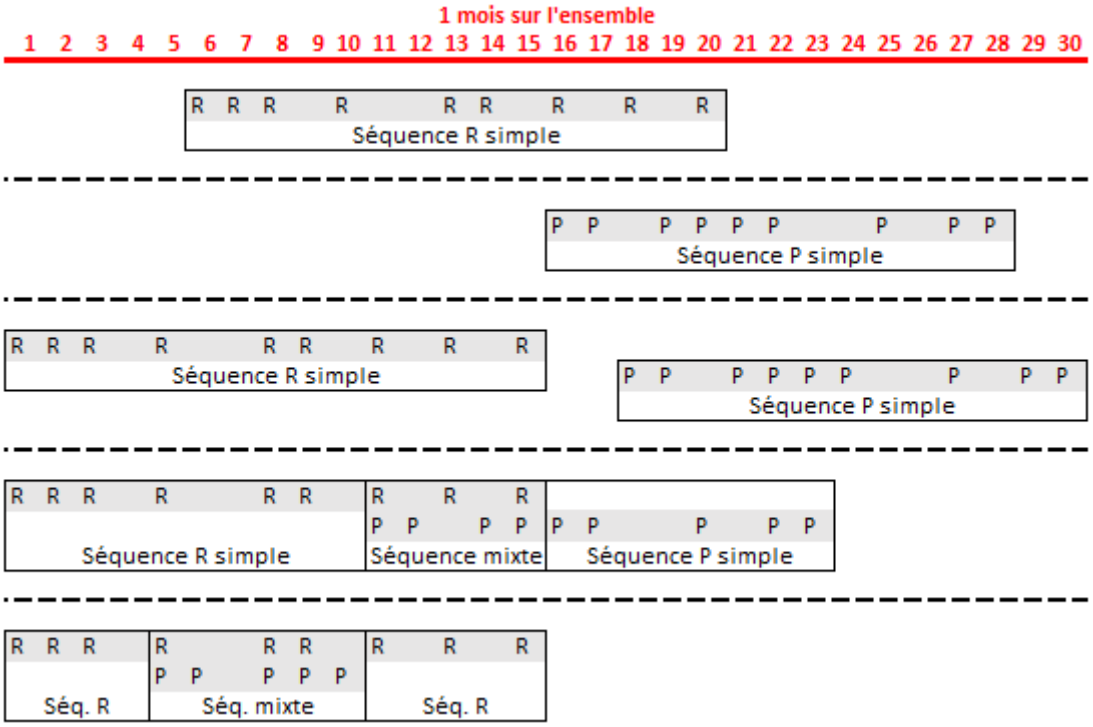
Le Stack Exchange Data Dump

Il s'agit d'une base de données très volumineuse mise à disposition sur archive.org, sous licence libre. La base est une copie brute anonymisée et exhaustive (plus de 300 Go) de l'ensemble des échanges ayant eu lieu sur la plateforme, elle est mise à disposition du public sous licence libre. Afin d'exploiter notre terrain d'étude, nous avons extrait, transformé et chargé les données brutes produites entre 2008 et 2021 en nous focalisant sur les publications étiquetées « R » et « Pandas » (Morge 2023). Cette archive est cependant menacée par les LLM (Large Language Models) qui l'utilisent pour alimenter des « IA » tels que ChatGPT*, lesquels détournent le public habituel de SO. Elle a brièvement cessé d'être rafraîchie en 2023 pour éviter son siphonage continuel. L'archive est constituée de différentes tables, téléchargeables séparément au format relationnel, c'est-à-dire livrées avec un identifiant anonyme permettant de réaliser des fusions. L'ensemble utilisé pour cette recherche s'est avéré trop lourd pour être géré, au moins initialement, sur des PC traditionnels. La plateforme de traitement massif ULille a donc été mise à contribution. Dans le cadre de cet article les tables mobilisées ont été au nombre de trois. La table CONTENTS contient le tag pour chaque publication sur StackOverflow, c'est-à-dire le thème sur lequel porte la question ou la réponse. Dans cette recherche nous nous sommes limités aux tags mentionnant les outils R ou Pandas (Python). La table CONTRIBUTES donne pour chaque publication un horodatage, l'identifiant propre et l'identifiant de la question s'il s'agit d'une réponse. La table USER fournit des informations sur les personnes, bien que leur identifiant principal soit dument anonymisés, certains codeurs restent identifiables légalement parce qu'ils délivrent volontairement et publiquement leurs identité ou site web personnels. La licence d'utilisation de l'archive impose même de les laisser apparaître tels que s'ils sont mentionnés dans une utilisation de la table. C'est dire la logique auto-promotionnelle inhérente à cette plateforme.*

https://archive.org/details/stackexchange

Pour avancer nous allons commencer par modifier le format de donnée pour rassembler les questions réponses dans des observations plus agrégées. Le terme de « séquence d'adhésion » désignera la période durant laquelle l'utilisateur (informaticien, *data analyst*, chercheur) est engagé dans un cumul de connaissances centré sur un des deux logiciels. Les « séquences d'adhésion mixtes » désignent des phases transitoires ou complètes, durant lesquelles un utilisateur peut alterner entre les deux langages sans totalement abandonner l'un au profit de l'autre. Avec des utilisateurs pouvant cumuler plusieurs centaines ou milliers de contributions sur des années, notre méthode va consister à reconstituer ces épisodes, et ce faisant à simplifier la base. Au final, chaque contributeur aura entre une (minimum) et trois (maximum, i.e. deux simples et une mixte) observations dans la source, en suivant le schéma suivant (schéma 1).

Schéma 1 : méthodologie de construction des épisodes, typologie d'usagers



Lecture : on donne ici l'exemple de 5 types de trajectoires possibles d'usagers sur la plateforme (cf. infra)

- *« R » : publication unique portant sur R*
- *« P » : publication unique portant sur Pandas*

L'enchaînement des publications forme des séquences (cf. tableau 2), l'enchaînement des séquence forme des trajectoires (cf. tableau 3)

Nous allons nous baser sur les bornes temporelles de chaque période durant laquelle un utilisateur est impliqué dans un des deux outils. Dans un premier temps on ne fait pas de discrimination entre les contributions de type question ou réponse. Ce que l'on distingue en revanche, c'est si la contribution se fait sur le logiciel R ou sur Pandas. L'approche va nous donner cinq types de séquences possibles. On a représenté ci-dessus cinq exemples fictifs sur une échelle temporelle d'un mois.

Les deux premiers types sont des séquences simples durant lesquelles un utilisateur n'émet que sur l'un des deux langages. Dans le troisième type l'usager transite brutalement depuis un langage vers le second sans revenir au premier. Dans le quatrième type un épisode de transition s'intercale durant lequel le contributeur poste à propos des deux langages alternativement, jusqu'à revenir à l'un des deux uniquement. Dans le cinquième et dernier type l'utilisateur connaît une période similaire mais revient à son premier engagement[5].

[5] Pour le moment nous ne laissons pas apparaître 3 autres types possibles : ceux entièrement mixte, ou commençant ou se terminant pas une période mixte. Un premier ou un dernier message posté inaugure en effet toujours une période R ou Pandas, mais pas les deux à la fois. Cette règle pourrait être assouplie par la suite.

On obtient une table de 107 320 épisodes R, Pandas ou Mixtes (cf. tableau 2, en remplacement des 1.2 millions d'observations initiales). Chaque contributeur est représenté par une (une seule séquence simple), deux (deux séquences simples à la suite sans chevauchement), ou trois lignes (deux séquences simples et une mixte entre les deux). Le nombre de contributeurs conservés est de 99 308 (l'autre moitié de notre cohorte n'a posté qu'un unique message, nous faisons le choix de les sortir des analyses pour le moment, ils réapparaîtront plus loin dans nos analyses).

Tableau 2 : conversion de la source en « séquences d'adhésion »

Type de séquence	N	%	Total Contrib	Moy. Contrib	Med. Contrib
R	63 045	58,7%	717 858	11,4	3
P	41 071	38,3%	332 137	8,1	3
Mixte	3 204	3,0%	99 308	31	6
Total	**107 320**	**100,0%**	1152508	107	3

Source : *Stack Exchange Data Dump*

On retrouve plus ou moins les mêmes proportions (entre R et Pandas) que dans le tableau 1. Le passage des contributions brutes aux séquences réduit un peu la représentation de R à 58,7%, ce qui est probablement un effet mécanique : le surcroît de contributions a tendance à être absorbé dans des séquences plus longues (ce que confirme la moyenne du nombre de contribution par séquence : 11,4 pour R, contre 8,1 pour Pandas). On commence aussi par relever un chiffre important : *les contributions mixtes représentent seulement 3% de l'ensemble*. L'interprétation peut être double : soit peu de contributeurs se convertissent d'un domaine à un autre, soit le font-ils de façon radicale, en cessant complètement de poster sur le domaine abandonné. On constate en revanche que ces séquences peu nombreuses sont beaucoup plus prolixes, avec une moyenne de contributions par séquence de 31, soit un chiffre bien supérieur aux domaines simples : il faut probablement y chercher un effet du profil des contributeurs.

Tableau 3 : trajectoires de contributeurs

		N	en % de l'ensemble	en % de la catégorie (séq. simples / multiples)
Séquences simples	Séquences simples sur R	57 702	58,3%	60,4%
	Séquences simples sur Pandas	35 994	36,4%	37,7%
Séquences simples avec incursions	R --> Mixte --> R	960	1,0%	1,0%
	P --> Mixte --> P	827	0,8%	0,9%
	Total	**95 483**	**96,5%**	**100,0%**
Séquences multiples	Migration brusque R vers Pandas	1 582	1,6%	46,2%
	avec transition (épisode mixte)	1 083	1,1%	31,6%
	Migration brusque Pandas vers R	424	0,4%	12,4%
	avec transition (épisode mixte)	334	0,3%	9,8%
	Total	**3 423**	**3,5%**	**100,0%**

Ensemble	**98 906**	**100,0%**	**-**

Source : *Stack Exchange Data Dump*

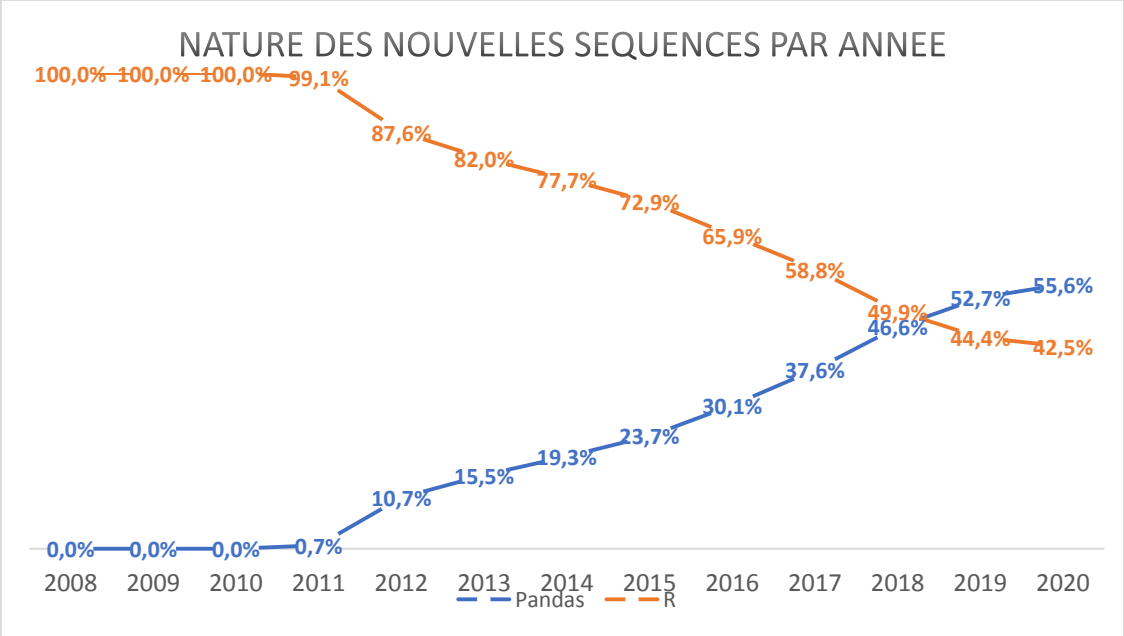
Le tableau 3 passe du niveau séquence au niveau individu (un même contributeur peut rassembler jusque trois séquences maximum). Le tableau continue de montrer l'importance de R, surtout dans sa partie supérieure. Environ 60% de l'ensemble des contributeurs sont mono-séquence et centrés sur R. Pandas représente un peu moins de 40% de l'ensemble. Ce résultat est normal, avec près de 10 ans d'avance pour R, le cumul est plus ancien. Ce qui est important pour nous se situe dans la deuxième partie du tableau :

- *(i) on constate que seuls 3.5% des contributeurs ont fait une conversion de l'un vers l'autre de ces logiciels (ligne total, séquences multiples) ;*
- *(ii) si Pandas ne représente que 36% des séquences, il absorbe en revanche près de 78% des conversions de l'un vers l'autre (1582 + 1083 sur 3423).*

Ce double constat est important. Les 3.5% montrent d'abord que les conversions sont rares : les contributeurs installés dans un logiciel donné ne vont pas massivement vers la nouveauté que constitue Pandas, un équivalent de R dont l'intérêt est d'être intégré dans un logiciel plus polyvalent (Python). En revanche la dynamique des conversions est à 80% environ en faveur de Pandas.

Cette approche synchronique ne fait pas apparaître la dynamique des communautés. Dans le graphique 1, on observe le volume d'ouvertures de nouvelles séquences selon les années et selon le type de séquence. Une nouvelle séquence appartient à deux types : R, ou Pandas. C'est une période temporelle durant laquelle un contributeur a consacré l'intégralité de ses messages à l'un ou l'autre des deux langages.

Graphique 1 : nouvelles séquences amorcées par années selon le type



Source : *Stack Exchange Data Dump*

Le graphique 1 montre nettement l'évolution de la communauté en faveur du nouveau logiciel par rapport à l'ancien. La croissance de la part des nouvelles séquences 100% Pandas commence dès

l'apparition du logiciel et augmente à un rythme soutenu jusqu'au croisement des deux courbes en 2018, moment de bascule où les néo-utilisateurs Pandas deviennent plus nombreux que ceux venant avec R.

Ces premiers résultats nous montrent d'une part qu'un changement est à l'œuvre à partir de 2011, c'est-à-dire dès la publication de Pandas. Le basculement en faveur de ce dernier se fait progressivement jusqu'à ce que les deux courbes se croisent en 2018. On assiste donc bien à une conversion épistémique d'un savoir vers un autre, dans une communauté de *data analysts* importante, sur une plateforme connue pour être l'alpha et l'oméga en matière de savoir-faire informatique. D'autre part, ces résultats montrent que cette transition ne s'opère pas ou très peu, en réalité, à l'échelle biographique individuelle. Les spécialistes de la data ne transitent pas facilement d'un logiciel à l'autre. Le changement serait intergénérationnel et viendrait donc plutôt d'un processus écologique de remplacement d'une masse d'individus par une autre.

Il ne s'agit en aucun cas de dire que le logiciel R entre en déshérence au cours de la décennie. Sur ce point la limite de notre approche est de focaliser uniquement sur les contributeurs actifs, questions posées ou réponses produites. A cela manquent deux aspects. D'abord le nombre de questions possibles sur un sujet donné va diminue nécessairement avec le temps, à mesure que des réponses leur sont trouvées, et que celles-ci sont stockées en rendues visible au grand public sur *SO*, qui n'est autre qu'une vaste archive de questions répondues (une FAQ exhaustive). Le déclin des publications concernant R pourrait ainsi simplement traduire l'avènement d'un savoir constitué, dont la FAQ est arrivée à maturité, mais qui continue d'alimenter un volume important de connexion de la part de consommateurs purs (c'est-à-dire de personnes ne posant ni ne répondant à aucun problème lié au logiciel), ne faisant que consulter la FAQ. Cette limite possible de notre approche est elle-même à relativiser, puisqu'on peut supposer qu'un logiciel qui cesse son évolution entre tout de même dans une forme de déclin (parce qu'il suscite moins de questions, parce qu'il cesse les améliorations progressives apportées aux réponses possibles, etc.[6]).

Les *Threshold Models* élaborés en SNA ne s'appliquent donc pas correctement pour décrire cette situation. On est bien face à une forme d'action collective (l'adoption d'une nouvelle pratique par une communauté professionnelles) mais, en premier lieu, cette action se pose en rupture à l'existant, en intervenant dans un espace structuré préalablement. En second lieu les nouveaux acteurs qui y prennent part intègrent le réseau depuis l'extérieur, et y apportent leurs nouvelles pratiques. Le basculement de la communauté est exogène plutôt qu'endogène. Si le premier constat a bien été pris en compte dans la littérature (par exemple par Übler et Hartmann 2016, qui modélisent bien des tendances nouvelles, avec une réflexion sur la défection des acteurs par rapport aux anciennes tendances), les modèles en analyse de réseaux ne portent pas ou peu sur le changement de normes s'opérant lors du renouvellement des générations et/ou des membres d'un groupe.

Du point de vue de l'analyse de réseaux nous avançons donc deux résultats importants. D'abord il y a bien l'action d'un groupe social d'acteurs qui communiquent et qui amènent le changement. Ces communications interviennent sur un support entièrement numérique et se fondent exclusivement sur des échanges inter-individuels, caractérisant bien un réseau d'échange d'informations et de conseils, justiciable des théories et méthodes de l'analyse de réseaux sociaux. Ensuite ce groupe social, et surtout ce qu'il produit, a pour caractéristique un format éminemment intergénérationnel, c'est-à-dire avec un niveau important de renouvellement des nœuds du réseau. Des acteurs partent avec leurs anciennes pratiques (ou se rendent moins visibles), d'autres y entrent avec de nouvelles

[6] Les critiques faites à R reposent aussi beaucoup sur le cumul désordonné des couches successives, là où un système comme Python repose sur un assainissement plus régulier de ses bases.

habitudes de travail. Et le changement s'opère sans basculement normatif de l'acteur individuel, sans révolution de ses propres normes ou de son ontologie, ce que s'évertuent pourtant à tenter de modéliser la plupart des recherches que nous avons citées dans le domaine de la SNA (i.e. le « seuil de tolérance » au changement de l'acteur et ses effets sur le groupe). Sur un sujet pourtant aussi peu sensible qu'un changement d'outil logiciel ou de méthodologie, tel que nous l'étudions ici, l'acteur individuel résiste : peu font la transition, et pourtant en une courte décennie, tout a changé. S'il en est ainsi pour un simple changement de *software*, que devons-nous penser des opinions morales, politiques ou esthétiques ?

Dans la suite de cet article nous proposons d'aborder le phénomène sous deux angles, à commencer par les mécanismes de la scission, qui explorent plus avant la dynamique de l'apparition d'une hétérodoxie sur ces forums (en l'occurrence, Pandas). Traditionnellement en sociologie, l'adhésion aux pratiques, aux valeurs, normes, etc., se fait à travers le phénomène de la socialisation. Plus spécifiquement le respect des normes épistémiques, dans un groupe de techniciens, ingénieurs ou chercheurs comme on s'y intéresse ici, joue un rôle essentiel dans sa capacité de coordination (Henderson et Graham 2017), son non-respect peut engendrer des externalités d'adoption négatives, c'est-à-dire l'incapacité de l'acteur individuel à continuer à travailler efficacement avec son entourage (Galeotti, Goyal 2009). On doit donc se demander pourquoi et comment ce mécanisme fondamental n'existe pas ou est neutralisé sur la décennie que nous étudions, à travers une étude du « précursorat antagoniste » en informatique : autrement dit, pourquoi un groupe d'acteur décide-t-il subitement de changer la norme dominante, et qui sont ces acteurs (en SNA cette question est nommée *start-up problem*, cf. Macy et Evtushenko 2020) ? Et enfin comment le réseau suit ou s'adapte, s'il le fait seulement ?

Scission méthodologique minoritaire sur StackOverflow

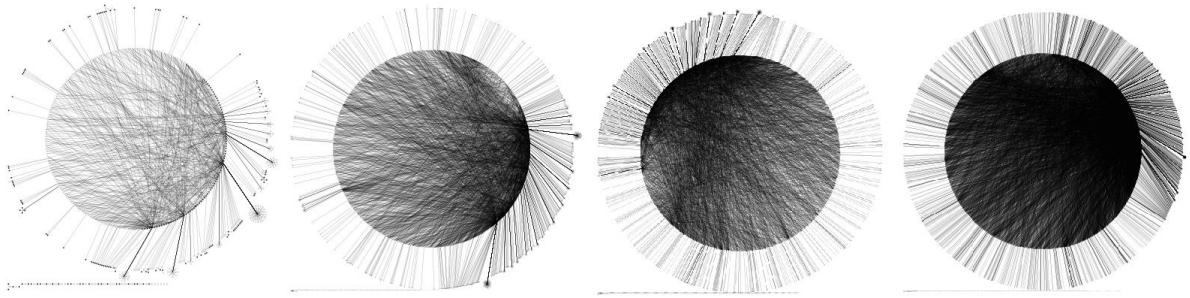
L'objet de cette partie est d'analyser les conditions dans lesquelles des acteurs minoritaires agissent à contre-courant d'une logique dominante. Dès les prémisses des modèles de réseaux sur les effets de masse critique, le phénomène a été codifié en tant que *Snob Effect* ou *Reverse Bandwagon*. Dans le cas d'une approche basée sur la consommation de biens et services, avec pour exemple d'interprétation les effets de snobisme dans la mode notamment : un groupe d'acteur ne bascule pas dans la pratique dominante précisément parce que la masse le fait, et il fomente la pratique suivante (Granovetter et Soong 1986). Dans le cas de SO que nous étudions, sur la période 2010-2012, on voit en effet la naissance d'un groupe d'utilisateurs de Pandas, le concurrent de R, alors que ce dernier connait une forme d'apogée avec près de dix ans d'existence et des dizaines de milliers de contributeurs sur la plateforme. Dans ce cas, l'effet de snobisme nous paraît difficilement applicable (même s'il existe certainement des langages « snobs » en informatique, chaque champ ayant ses élites auto-proclamées).

Nous allons donc zoomer sur cette scission minoritaire, laquelle deviendra dominante à partir de 2018. Nous constatons en effet que certains acteurs fortement minoritaires amorcent un phénomène de pompe et propulsent un logiciel nouveau face à un autre, pourtant déjà bien installé. Les données de SO délivrent quelques indices permettant de caractériser ces individus, et le fonctionnement de leur collectif. C'est ce que nous allons voir dans la suite de ce texte, consacré aux

prémisses, c'est-à-dire à cette période de rupture durant laquelle ce sous-groupe de néophiles joue un rôle caractéristique.

Commençons par observer sous la forme de graphes, les quatre premières années du développement de la communauté Pandas (graphe 1). Ces réseaux montrent que le cumul d'échanges de contributions autour de Pandas est déjà conséquent vers fin 2012, à la sortie du logiciel. La densification du groupe va se poursuivre à un rythme très rapide les années suivantes. Au premier trimestre 2012, moins de 50 personnes échangent sur Python, dont 28 ne font que questionner, et 18 apportent des solutions. Au second trimestre de la même année on observe un quasi triplement des utilisateurs s'informant ou informant sur Pandas. Le nombre d'individus répondant à des questions passe de 18 à 47. Fin 2012 les échanges sur Pandas concernent ainsi 562 personnes, dont 38% de personnes répondant à au moins un problème soulevé par une autre (première ellipse du graphe 1). En 2013 ces valeurs connaissent un triplement, avec toujours un taux d'« encadrement » (proportion d'individus qui répondent aux questions posées sur le total des individus contributeurs) qui reste stable (37%). En 2014, la communauté double (39% pour le taux d'encadrement), et fin 2015 on arrive à environ 6000 utilisateurs, dont près de 40% répondent principalement aux questions des autres.

Graphe 1 : évolution de la communauté Pandas lors de ses prémisses (2012-2015)



Source : Stack Exchange Data Dump

Lecture : chaque sommet représente un utilisateur de SO, chaque lien indique une réponse apportée à une question

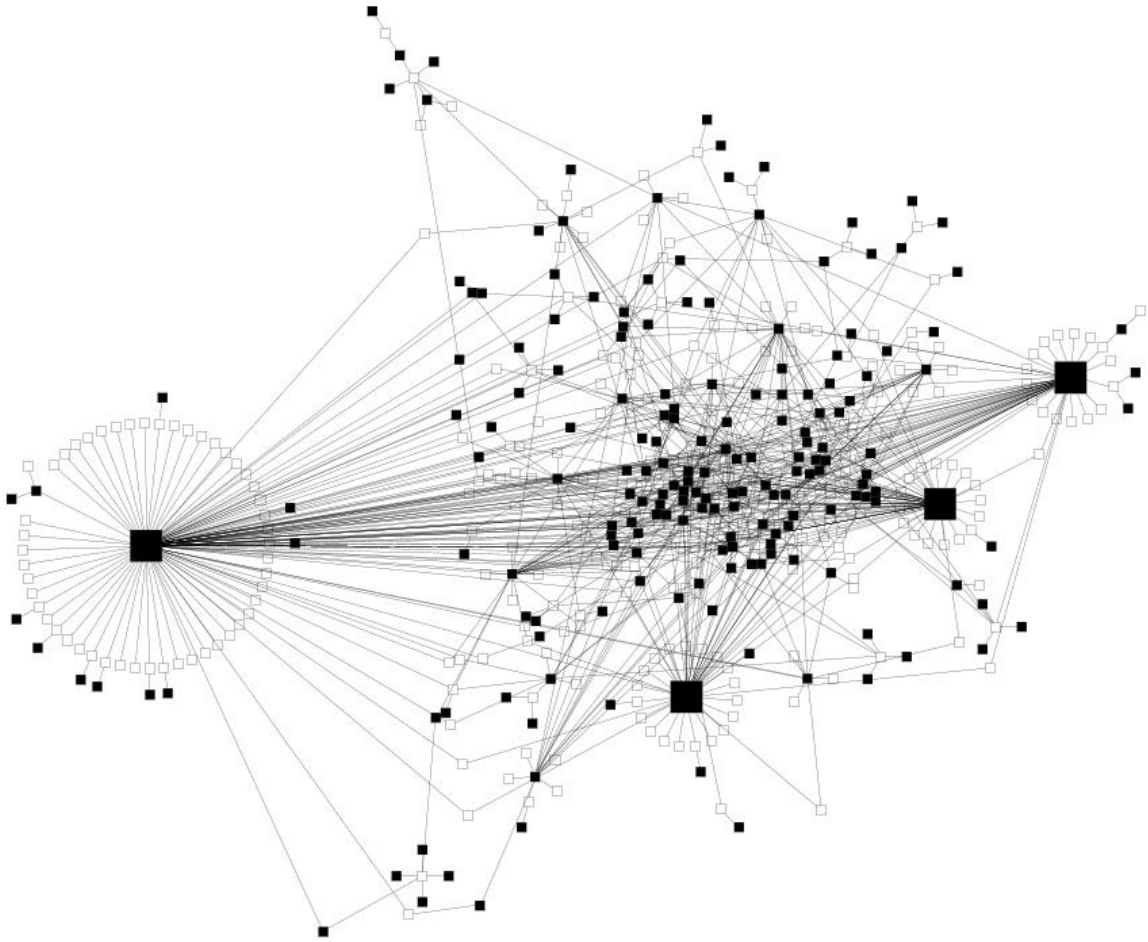
La croissance importante de la communauté commence donc dès 2012 (graphe 1). Un zoom sur cette année essentielle (cf. graphe 2) fait apparaître le rôle d'une personne très active dès le début de l'année (à gauche du graphe 2), laquelle sera rejointe en milieu d'année par trois autres (à droite en descendant). Ces individus ne communiquent pas entre eux et semblent s'ignorer réciproquement (loin s'en faut, nous allons le voir dans la suite). Par ailleurs ils ne posent aucune question, mais répondent à une masse importante de demandes, et ce dans la période cruciale qui suit le lancement du package Pandas.

Or l'étude des pseudonymes va nous permettre de mieux comprendre les forces agissantes. En croisant les données de l'archive SO (qui donne les pseudonymess, puisqu'ils sont publics) avec celles du site Web actif, on commence à mieux comprendre d'où vient la scission qui nous intéresse. La personne très active dès le début de l'année (à gauche du graphe 2) est Wes McKinney, présent toute l'année, répondant à la plupart des questions sur Pandas. Il n'est autre que le créateur du package Pandas lui-même (« *benevolent dictator for life* », comme indiqué sur la page *Github* du

logiciel). McKinney est également l'auteur de manuels traduits et très diffusés aux éditions O'Reilly, supports bien connus des enseignants universitaires, dont les premiers sortent en… 2012 ((McKinney 2012) mais aussi un peu avant au format *working paper* McKinney 2011). Ce graphe de l'année 2012 permet ensuite de repérer trois autres acteurs, plus mineurs, mais aussi très actifs à partir du milieu de l'année (contrairement à Mc Kinney, présent dès le début de l'année). Chang She en premier lieu, est un ancien camarade de classe du MIT de Mc Kinney. Il figure sur la liste des vingt membres de l'équipe de direction de Pandas. Associé en affaire, il a cofondé avec Mc Kinney un logiciel de type tableau de bord, *Datapad.* Les deux autres sommets importants visibles sur le graphe représentent un ingénieur Belge travaillant chez *Broadcom*, Wouter Overmeire, et un data-analyst américain freelance, Andy Hayden. Sans surprise ces deux spécialistes sont également sur la liste des vingt administrateurs de Pandas (le premier est cependant listé comme ancien membre).

On se rend donc facilement compte que la pompe (le « *start-up problem* » dans les modèles en SNA) est amorcée par les créateurs du logiciel eux-mêmes. L'existence sur ces plateformes de telles *task-forces* organisées en *back office* ne sont pas rares : comme ici elles rassemblent souvent de moins de dix personnes, et il s'agit d'un schéma promotionnel commercialement rodé (cf. l'analyse sur *GitHub* et *StackOverflow* faite par Sun et al. 2018). Ces personnes intéressées donnent de leur temps sur *Stack* pour assurer le lancement du logiciel sur la plateforme : c'est ce qui s'observe très nettement en 2012.

Graphe 2 : la communauté Pandas naissante en 2012



Source : *Stack Exchange Data Dump*

Reste à élucider ce qui se passe ensuite, dans la mesure où le rôle de ces acteurs intéressés va se dissoudre dans une masse et des mécaniques différentes. L'études des « bio » et des pseudonymes des acteurs n'est d'ailleurs plus possible, à cause du nombre d'abord, mais aussi parce que ces utilisateurs plus secondaires ne mettent pas systématiquement en avant leurs c.v. ou identités, pour des raisons évidentes de statut professionnel. En effet, si les créateurs de Pandas sont des indépendants reconnus et médiatiques, les informaticiens salariés n'ont peut-être pas intérêt à ce que leur employeur réalise qu'ils passent un temps parfois conséquent à accumuler du capital réputationnel sur SO en parallèle de leurs missions salariées (nous reviendrons sur ce point plus loin).

Le passage d'un régime basé sur le support d'un faible nombre d'acteurs ayant un intérêt évident à ce que l'utilisation du logiciel se répande, à un régime basé sur un soutien de masse anonyme et public, fait l'objet des analyses qui suivent. Il s'agit de voir comment ces *insiders* parviennent à amorcer la pompe et à déclencher un intérêt suffisant de la part du public pour que le volume de questions / réponses sur SO atteignent une masse critique, puisqu'ils ne peuvent pas, matériellement, répondre à tous[7]. Pour cela nous allons reprendre l'approche par séquences développée en première partie, et plutôt que de qualifier des épisodes R, Pandas ou alternant entre les deux, nous allons voir la façon dont les contributeurs sur SO transitent ou non du statut de débutant interrogeant à la cantonade, à celui d'expert érudit répondant aux questions des autres. La dynamique de ces deux statuts devrait permettre de caractériser la réussite et la massification de l'utilisation du package Pandas.

Précurseurs, comptes étoilés, et masses anonymes dans le développement de la Super-FAQ Pandas

Analyser le succès d'un oligopole de fondateurs autour de Pandas signifie reconnaître les divisions internes de cette communauté *post-FLOSS (Free Libre Open-Source Software)*, en dépit de son idéologie égalitaire et libre. Le logiciel ne peut pas démarrer sans un groupe d'individus dont le niveau de savoir est anormalement élevé par rapport à un second groupe, beaucoup plus massif, d'utilisateurs qui soumettent à ces experts les problématiques qu'ils rencontrent, dans une diversité d'usages que les premiers ne peuvent pas anticiper au départ, et qu'ils souhaitent accompagner. Pour ce faire nous allons modifier notre approche des séquences telles que présentées en premières partie.

L'inconvénient dans la conception des séquences proposées dans le schéma 1 est qu'elle nécessite pour chaque séquence *a minima* deux événements, l'un pour l'ouverture et l'autre pour la fermeture de la séquence. Nous nommerons ces séquences « normales » parce qu'elles traduisent un

[7] Chakraborty et al. 2021 posent le même constat sur le rôle de promotion assuré par les développeurs en début de dynamique, toujours sur StackOverflow, mais concernant les langages *Swift*, *Go* et *Rust* ; Galeotti et Goyal 2009, 2010 approchent le problème à partir de la modélisation de la façon dont des influenceurs s'y prennent pour favoriser l'émergence d'un nouveau produit dans un réseau donné, Macy 1991 soulève aussi le rôle d'une minorité d'acteurs puissants et intéressés

comportement des contributeurs qui correspond à une activité régulière minimale justifiant la création d'un compte sur la plateforme. Nous allons voir que de nombreuses entrées sur SO ne correspondent pas à ces séquences normales, et que de nombreux membres inscrits ne postent qu'une fois ou deux, mais faisant masse, y jouent tout de même un rôle essentiel que ne sauraient remplir les seuls utilisateurs réguliers. Cette vision dézoomée sur notre jeu de données a pour effet de réintégrer un volume important d'acteurs, comme on pourra le voir dans le tableau 5.

La seconde transformation que nous opérons sur l'approche séquence est de les former non plus sur la base d'un langage particulier (R *versus* Pandas), mais sur le rôle principal joué par chaque utilisateur : pose-t-il des questions, répond-il à des questions, ou fait-il les deux ? Nous aurons donc des séquences d'utilisateurs « questionnants » ou « répondants », un même contributeur « alternant » pouvant passer de l'une à l'autre, et des séquences mixtes durant lesquelles questions et réponses s'enchaînent sans schéma dominant.

Tableau 5 : classification empirique des séquences questions / réponses

	Pandas		R		Ensemble	
Mono pur dont :						
mono pur questionnant	*39139*	*35,8%*	*58400*	*35,4%*	*90591*	*34,6%*
mono pur répondant	*19894*	*18,2%*	*18737*	*11,4%*	*33935*	*13,0%*
Séquences normales dont :						
séquences questionnant	*23322*	*21,3%*	46440	*28,2%*	*69653*	*26,6%*
séquences répondant	*13189*	*12,1%*	15643	*9,5%*	*28223*	*10,8%*
séquences mixtes	*3321*	*3,0%*	9678	*5,9%*	*13523*	*5,2%*
Monos mixtes	4418	*4,0%*	5718	*3,5%*	9367	*3,6%*
Mono intra	3234	*3,0%*	5773	*3,5%*	8988	*3,4%*
Mono post	1805	*1,7%*	2983	*1,8%*	4787	*1,8%*
Mono ante	1055	*1,0%*	1531	*0,9%*	2590	*1,0%*
Total	109377	*100,0%*	164903	*100,0%*	261657	*100,0%*

Source : *Stack Exchange Data Dump*

Lecture :

- « mono pur questionnant » : une seule question posée par un seul usager

- « mono pur répondant » : une seule réponse donnée par un seul usager

- « séquence questionnant » : une série de questions posées par un seul usager

- « séquence répondant » : une série de réponses apportées par un seul usager

- « séquence mixtes » : alternance de question(s) et de réponse(s)

- « monos mixtes » : une seule question et une seule réponse uniques, dans n'importe quel ordre

- « mono intra » : une question isolée au milieu d'une série de réponses ou le contraire

- « mono post » : une question isolée après une série de réponses ou le contraire

- « mono ante » : une question isolée avant une série de réponses ou le contraire

Dans la dernière colonne, on retrouve plus ou moins le même total de séquences normales que dans le tableau 2 (110 000 environ), ce qui est une première indication de la faible mobilité des utilisateurs entre les rôles, nous y reviendrons plus loin. Et en dehors de ces séquences normales (*a minima* deux entrées, soit sur R, soit sur Pandas), nous faisons deux observations simples : la réintégration des utilisateurs ayant été écartés dans la première phase de notre démarche fait plus que doubler le volume des cas pris en compte, passant à plus de 260 000 (dans les mêmes proportions, là encore, que la différence observable dans le total de contributions entre les tableaux 1 et 2).

Cette masse anonyme d'utilisateurs sans rôles individuels identifiés pèse donc faiblement dès lors qu'il s'agit de regarder le total des contributions individuelles, mais pèse lourd si on considère les séquences, qui sont techniquement la première façon de distinguer des de rôles sur la plateforme. *Nous observons donc que des individus sans rôle pèsent davantage que les autres, leur va-et-vient numérique permanent engendre une masse de questions qui alimente la plateforme plus que ne le font les contributions régulières des individus « à rôle »* c'est-à-dire ayant un compte justiciable des règles d'évaluation de la plateforme (badges, notations, classements, etc.), et/ou renvoyant explicitement à un profil professionnel qui lève son anonymat à dessein (site personnel, profil linkedIn).

Au passage, on observe que les écarts entre Pandas et R ne sont pas manifestes. Les rôles se distribuent plus ou moins à l'identique, à savoir une masse anonyme très majoritaires de néophytes, et un groupe plus réduit d'experts, sur lesquels nous allons zoomer dans la suite (en nous basant sur Pandas uniquement), et dont nous avons la biographie complète grâce à l'archive (R étant plus ancien, nos données ne couvrent pas ses premières années d'existence).

Le tableau 6 dresse ensuite le portrait de l'ensemble Pandas, sous la forme d'une classification finale des utilisateurs. Comme dans la première partie de ce document, nous combinons les séquences brutes visibles dans le tableau 5 pour observer la façon dont un acteur passe d'un rôle à un autre (en première partie nous le faisions pour le passage d'un logiciel à un autre). Nous cherchons à reconnaître des rôles plus élaborés aux individus, en fonction de leur comportement sur la plateforme.

On observe d'abord qu'une majorité des entrées sur la plateforme renvoie à des incursions brèves d'*utilisateurs sans statut*. Ils représentent plus de 60% de l'ensemble, soit une large majorité. Il s'agit d'individus ayant créé un compte sur *StackOverflow* et ne l'ayant utilisé qu'une seule fois, la plupart du temps pour ne poser qu'une seule question.

Ensuite, on peut distinguer la catégorie des *statutaires de second ordre*, qui rassemble les contributeurs ayant des profils centrés sur la demande d'aide. Ils comptent pour un peu moins d'un quart des inscrits sur SO. Ils se distinguent des premiers par le fait qu'ils postent une série de questions plutôt qu'une seule. Ils jouent aussi un rôle plus important de par le volume des entrées qu'ils déposent dans la FAQ, mais aussi probablement par le niveau de technicité supérieur de leurs interventions, à mesure qu'ils gagnent eux-mêmes en expérience, trouvant des réponses, et raffinant leurs questions. En cumulés, les *utilisateurs sans statut* et les *statutaires de second ordre* représentent plus de 85% des contributeurs de la plateforme StackOverlflow.

Les deux catégories restantes comprennent donc les contributeurs experts, ceux qui alimentent le site en savoir-faires, répondent aux questions à la volée, assistent autrui, dans un rapport très déséquilibré de un pour neuf. Sans eux, pas de communauté en ligne. Cette catégorie se décompose en deux. Premièrement elle comprend les *statutaires ambivalents*, dont le rôle est mixte, ils sont très rares et n'englobent que 3% du total. Il s'agit pour l'essentiel de *statutaires ambivalents en mobilité ascendante*, parce qu'ils passent du rôle d'usager interrogeant les autres, à celui d'expert

assistant. La rareté de ce profil témoigne d'une forte viscosité sociale, autrement dit d'une absence de mobilité ascendante sur cette plateforme. On entre dans un rôle, on y reste. Deuxièmement, la dernière catégorie regroupe les individus jouant un rôle central sur SO, à savoir les *statutaires de premier ordre*, c'est-à-dire ceux qui ne font que répondre aux questions, et qui le font régulièrement. Ils représentent un usager sur dix. Parmi eux rares sont ceux qui se permettent ne serait-ce qu'une seule question sur le forum, ils campent exclusivement dans leur position d'expert.

Tableau 6 : classification finale des rôles sur la plateforme

Utilisateurs sans statut		Usager ne posant qu'une seule question	39139	39,4%
		Usager ne publiant qu'une seule réponse	19894	20,0%
		Usager posant une question suivie d'une réponse	3815	3,8%
		Usager apportant une réponse suivie d'une question	603	0,6%
Statutaires de second ordre		Usager posant une série de questions (séquence simple)	16989	17,1%
		Usager posant une série de questions et apportant une réponse isolée dans l'intervalle (séquence mixte)	2490	2,5%
		Usager posant une série de questions et apportant plusieurs réponses isolées dans l'intervalle (séquence mixte)	1301	1,3%
		Usager posant une série de questions et apportant une réponse isolée après l'intervalle (séquence mixte)	1463	1,5%
		Usager posant une série de question après avoir apporté une seule réponse (séquence mixte)	234	0,2%
Statutaires ambivalents	en mobilité descendante	Usager passant progressivement du statut de répondant à celui de questionneur(séquence mixte)	308	0,3%
		Usager passant du statut de répondant à celui de questionneur sans transition (séquence mixte)	106	0,1%
	en mobilité ascendante	Usager passant progressivement du statut de questionneur à celui de répondant (séquence mixte)	1217	1,2%
		Usager passant du statut de questionneur à celui de répondant sans transition (séquence mixte)	403	0,4%
Statutaires de premier ordre		Usager apportant une série de réponses (séquence simple)	9002	9,1%
		Usager apportant une série de réponses après avoir posé une seule question (séquence mixte)	821	0,8%
		Usager apportant une série de réponses et posant plusieurs questions isolées dans l'intervalle (séquence mixte)	495	0,5%
		Usager apportant une série de réponses et posant une question isolée dans l'intervalle (séquence mixte)	744	0,7%
		Usager apportant une série de réponses et posant une question isolée après l'intervalle (séquence mixte)	342	0,3%
Total			99366	100,0%

Source : *Stack Exchange Data Dump*

Synthétisé, ce tableau permet de poser trois constats essentiels :

- le « Q&A Site » *StackOverlfow*, le plus central en informatique sur la décennie étudiée, comprend beaucoup plus de personnes en demande de conseil que de personnes susceptibles de répondre à ces besoins, dans un rapport de un pour neuf environ ;

- la mobilité de statut entre rôle en est complètement absente : les personnes devenant ou s'estimant suffisamment qualifiées pour répondre aux questions des autres après avoir été elles-mêmes dans cette situation sont très rares. Il n'y a pas de mobilité sociale observable, les statuts sont fixes et pérennes, soit on pose des questions, soit on y répond, mais on ne fait pas les deux ;
- une minorité de contributeurs répond aux questions, et ne fait quasiment que cela (parmi ces statutaires de premier ordre, ceux n'ayant jamais posé de question sont huit fois plus nombreux que ceux interrogeant occasionnellement).

Ces différents points nous conduisent à une interprétation centrale : la source de savoir n'est pas inhérente mais bien extérieure à la plateforme. Les statutaires de premier ordre ne l'utilisant pas eux-mêmes pour développer leurs compétences, il est logique d'en déduire qu'ils se forment ailleurs, en dehors de celle-ci. Ainsi, paradoxalement, loin d'émerger d'une communauté horizontale dans laquelle les internautes se rendraient des services mutuels dans un vaste système d'échange généralisé, le savoir généré et mis à disposition sur cette super-FAQ dépend bien plus de mécanismes verticaux et exogènes. On souligne donc l'hétéronomie de la plateforme et, plus largement, la dépendance des communautés qui s'y forment vis-à-vis de stratégies communicationnelles et commerciales externes. De plus ce mécanisme quasi purement vertical n'est pas linéaire : les répondants répondent et ne questionnent pas eux-mêmes, on n'observe donc pas une forme de hiérarchie où des répondants d'un niveau d'expertise donné répondraient à des répondants de niveau d'expertise inférieur, jusqu'à la couche des questionnants purs. La plateforme est donc une dualité binaire plutôt qu'une hiérarchie.

On l'a vu plus haut (graphe 2), à l'origine (les prémisses), ces experts exclusivement répondants, dépositaires de ressources externes, ne sont autres que les fondateurs du logiciel. Ils ne sont qu'une poignée sur les six premiers mois (Wes Mc Kinney, Chang She, Wouter Overmeire, Andy Hayden). Viennent ensuite d'autres statutaires de premier ordre, une dizaine de milliers de personnes, sur lesquels nous allons tenter à présent de zoomer pour obtenir une caractérisation plus poussée.

Le tableau 7 reprend la typologie du tableau 6, afin d'analyser qui laisse des informations personnelles en fonction de son statut. Cette levée d'anonymat volontaire, possible sur le site, constitue un indicateur des finalités de la création du compte, et surtout explique les raisons de l'investissement (en volume de contributions) de certains individus sur la plateforme. Trois variables sont présentes, largement corrélées : une première est la présence d'un lien vers un site web personnel (colonne « Page web »), une deuxième est l'existence d'une biographie personnelle donnée sur la page d'accueil du compte QO (colonne « Bio »), enfin une troisième est l'indication d'une localisation, généralement le duo ville / pays (colonne « Localisation »).

Tableau 7 : informations personnelles laissées par les usagers en fonction de leur statut

	Page web		Bio		Localisation	
	oui	non	oui	non	oui	non
Utilisateurs sans statut	6325	57126	13267	50184	26456	36995
	10,0%	90,0%	20,9%	79,1%	41,7%	58,3%
Statutaires de second ordre	1487	20990	3998	18479	8474	14003
	6,6%	93,4%	17,8%	82,2%	37,7%	62,3%

Statutaires ambivalents	410 20,2%	1624 79,8%	926 45,5%	1108 54,5%	1193 58,7%	841 41,3%
Statutaires de premier ordre	2686 23,6%	8718 76,4%	5719 50,1%	5685 49,9%	7347 64,4%	4057 35,6%
dont les 5% les plus productifs	*157* *27,4%*	*416* *72,6%*	*348* *60,7%*	*225* *39,3%*	*414* *72,3%*	*159* *27,7%*
Total	**10908**	**88458**	**23910**	**75456**	**43470**	**55896**
	11,0%	**89,0%**	**24,1%**	**75,9%**	**43,7%**	**56,3%**
	99366					

Source : *Stack Exchange Data Dump*

Sur les six premiers mois qui ont suivi la création de Pandas, la FAQ s'alimente essentiellement à partir des contributions d'un quarteron de comptes n'étant autres que les informaticiens ayant lancé Pandas. Puis sur le moyen et long terme, se constitue ce groupe de 20 000 membres environ, dont le seul usage de SO consiste à répondre aux questions des autres utilisateurs, comme l'ont fait avant eux les créateurs du logiciel. Cet altruisme apparent n'est cependant pas toujours anonyme, et l'on peut supposer qu'il existe certaines logiques compensatoires en terme d'acquisition de réputation (tableau 7). On observe ainsi que les contributeurs sont d'autant plus prompts à sortir de l'anonymat qu'ils sont plus actifs dans la communauté. Un lien vers une courte biographie personnelle par exemple est délivrée chez la moitié des statutaires premier ordre mais seulement pour 20% des personnes ne contribuant qu'une fois (utilisateurs sans statut) ou ne faisant qu'interroger (statutaires de second ordre), qui sont encore moins nombreux à le faire. Apparemment les utilisateurs cachent d'autant plus facilement leur identité qu'ils ne font que poser des questions, évitant ainsi le risque d'afficher leur noviciat sur la place publique. On observe le même type d'écart pour le fait de pointer vers un site web personnel : 23.6% pour les premiers contre 5.7% pour les seconds. Ces écarts sont encore plus importants si on les mesure pour le sous-groupe des 5% des contributeurs les plus actifs dans le sous-groupe des statutaires de premier ordre : près d'un tiers d'entre eux fait un renvoi vers un site personnel, et plus de 60% ont une page biographique dûment renseignée (+ 10 points par rapport à la catégorie générale des statutaires de premier ordre). Les mêmes écarts s'observent pour l'indication d'une localisation géographique.

Les contributeurs importants à la plateforme, qui sont une minorité sur StackOverflow, sont donc aussi ceux qui ont le plus tendance à révéler leurs identités professionnelles. L'intérêt de SO est alors de servir de mesure de la compétence et de l'investissement, dans une logique de mise en avant de c.v. (même si tous ne le font peut-être pas aussi ouvertement, cf. supra). Parmi les contributeurs les plus actifs du groupe des statutaires de premier ordre, on retrouve des auteurs de manuels de programmation pour lesquels les réponses apportées sur SO sont un bon moyen de faire une promotion personnelle (Gordon Linoff ; Alvaro Fuentes par exemple), des développeurs *freelance* stars sur la plateforme (Martijn Pieters par exemple, dont la page web externe rappelle qu'il est le 7ème contributeur le plus prolifique de Stack, 1er s'agissant de Python, et qui enseigne aujourd'hui en visioconférence sur *CodeMentor* dans le cadre de sessions payantes de 15 minutes), des ingénieurs dans le domaine de la linguistique (Wiktor Stribiżew, dont le compte *linkedIn* cite également son classement sur *SO*), des statisticiens du secteur bancaire assez nombreux (tels Tim Biegeleisen ou John Zwinck par exemple), des statisticiens (Dirk Eddelbuettel), ou géologues (Joe Kington). Quelques chercheurs et ingénieurs français se trouvent dans ce groupe mais pour des volumes moins conséquents de réponses. Le simple fait de poster deux dizaines de réponses suffisant à les distinguer de la masse anonyme des individus qui ne font que questionner (Serge Ballesta, ingénieur

des Ponts à Météofrance ; Alvaro Fuentes, INRIA - Sophia Antipolis ; Ayoub Zarou, centralien, assurance / banque / finance ; Hugo le Moine *data scientist* dans le groupe de pharmacologie Merck ; Raphaele Adjerad, Ensae, administratrice Insee, etc.).

On l'a dit plus haut, StackOverflow est un système dual : une couche d'experts répondants qui ne font que cela, une couche de novices questionnants dans la même attitude. Il est à l'opposée d'un système d'échange généralisé à étages multiples où les plus compétents répondraient à d'autres moins connaisseurs, mais sollicitant tout de même de l'aide de temps à autre, et qui eux-mêmes aguilleraient l'étage immédiatement inférieur, etc. Ici pas de structure sociale en quasi hiérarchie, mais plutôt un marché de producteurs et de consommateurs, dans lequel la gratuité des réponses est inféodée aux externalités positives dont bénéficient certains des plus gros producteurs. Dans un tel système, il est clair que la viabilité d'un compte tiendra à sa capacité à se démarquer et à investir massivement les questions des utilisateurs sans statut, de façon à être en forte visibilité dans la super-FAQ, et à augmenter son score réputationnel, de même que les chances de voir des utilisateurs s'intéresser à sa page personnelle.

De nombreux sites web proposent des stratégies pour augmenter rapidement sa réputation sur StackOverflow (LinkedIn, Quora, le site de SO lui-même, et les pages personnelles des comptes importants sur la plateforme). De façon attendue, le principal conseil formulé consiste à répondre à un maximum de questions, en écrivant des réponses de qualité, si possible en première position (sur SO une question peut recevoir plusieurs réponses, celles-ci étant classées après coup à partir d'un système de votes). Dans un dispositif de ce type, le ratio entre le nombre d'utilisateurs posant des questions et le nombre de ceux apportant des réponses est crucial pour les seconds, dans la mesure où il va déterminer le potentiel de rente oligopolistique dont ces derniers peuvent jouir. Ces dépositaires de savoirs techniques exogènes ont donc intérêt à être actifs sur la plateforme tant que la niche n'est pas saturée de concurrents, parce qu'il s'agit d'une phase durant laquelle il est aisé d'accumuler la forme de capital numérique définie par les règles du jeu sur StackOverflow. A mesure que la niche intègre davantage de codeurs compétents attirés par le logiciel, la rente oligopolistique décroit, et la communauté change de régime en reposant sur des individus aux profils altruisme / opportunisme probablement plus équilibrés, qui fournissent des réponses plus étroites à des questions spécifiques (Chaoui et al. 2022; Delarre et al. 2023). Une part de la dynamique d'ensemble s'explique ainsi par des mécanismes similaires à ceux repérés en sociologie économique et des organisations : les entrepreneurs et entreprises à la recherche de « *non-competitive market niches* » (Fligstein 2002), de niches de qualité et sociales (White 2002, Lazega 2003), et de rentes de situations provisoires (Sørensen 2000).

Mais on observe ici que ces différents profils alternent, entrent et animent le réseau dynamiquement en le changeant de fond en comble, sans qu'il ne s'agisse jamais d'un ensemble fini d'acteurs se convertissant eux-mêmes, acceptant le changement et s'y adaptant. Dans cette perspective un rapport pour la DARES montrait bien les injonctions faites aux informaticiens s'agissant de leurs carrières : les « jeunes » (moins de 35 ans) sont spécialistes dans un savoir technique qu'ils offrent en tant que service, généralement dans une ESN (« Entreprise de services numériques », les anciennes SSII), avant de « passer en fixe » dans une société cliente, le plus tôt possible. Et la norme est que, passé l'âge de 35 ans, il faut s'être « *extirpé de la technique* » et passer du métier de développeur à celui d'encadrant, de conseiller ou de commercial, et donc abandonner la maîtrise technique fine du code informatique (Poussou-Plesse et al. 2008). Les informaticiens peuvent en effet difficilement se revendiquer de telle ou telle nouveauté logicielle, au-delà de quinze ans après leur formation diplômante. D'où probablement la quasi absence de conversions autres qu'intergénérationnelles observée sur SO.

Dans ce contexte, notre classification finale repère en définitive quatre groupes d'utilisateurs, répartis sur deux niveaux. Loin d'un système d'échange généralisé multi-couches. Les trois premiers groupes ne font que produire, et le dernier que consommer. En outre les trois premiers groupes ne parlent pas entre eux mais sont entièrement tournés vers la couche des utilisateurs « consommateurs ». Ce n'est donc pas à proprement parler d'un « réseau » qu'il s'agit, mais plutôt d'un système de production industriel à intégration verticale :

- les grands précurseurs intervenant dans un paysage logiciel très émergent et très peu dense (Wes Mc Kinney et ses associés), sont associés en externe au sujet dont ils sont les porteurs, dont ils font la promotion intéressée (en l'occurrence la bibliothèque Pandas devenu plus tard une ressource importante voire dominante en statistique). La plateforme StackOverflow leur offre un espace publicitaire considérable parce que l'information n'y dépend d'aucune structure sociale ou réseau, elle est offerte à tous (on n'y répond pas, en privé, à tel ou tel acteur, toute réponse est publique). Pour ces contributeurs, SO est un espace secondaire soumis à des stratégies exogènes. En suivant la typologie de Ragouet (Ragouet 2000) on dira que leur réputation est davantage « arborée » (dans sa conceptualisation de la notion de réputation, l'auteur oppose ainsi sa dimension verticale – importance de ego dans une arène – à sa dimension horizontale, par métaphore « arborée » – à savoir la diversité des arènes investies). Il est donc logique que ces précurseurs quittent le jeu plus librement que les autres, et qu'on ne les retrouve pas *in fine* dans la liste des utilisateurs les mieux évalués de la plateforme, puisqu'ils jouent sur une variété d'espaces, dont SO n'est qu'une phase locale et temporaire ;
- il en va différemment des contributeurs stars prenant leur suite (les « comptes étoilés » de SO). Sans intervenir à l'orée de la dynamique, ils vont y rentrer suffisamment tôt pour devenir de gros accumulateurs de capital numérique (doublant les précurseurs qui quittent le jeu une fois la pompe amorcée). Leur accumulation de réputation est plus spécifiquement interne. Mais en dominant l'arène, ces contributeurs stars peuvent ensuite assurer la conversion du capital numérique accumulé sur la plateforme, en publicisant leurs scores, leurs classements ou leurs médailles, comme on peut le voir sur leurs pages personnelles,
- il faut ensuite une grande masse d'utilisateurs plus anonymes, apportant de l'aide *gratis pro deo*, dans une volumétrie globale importante, mais trop insuffisante individuellement pour favoriser leur entrée dans le gratin des individus susceptibles de tirer un profit autre que symbolique de leur activité sur SO. Ils sont probablement ceux qui correspondent le plus à l'image de l'internaute altruiste respectant un certain esprit du libre et de l'entraide sur le web. Leur engagement de moyen ou long terme peut se comprendre sociologiquement à partir du crédit apporté à la compétence d'autrui, et leur souhait d'identification à lui (comptes suivis, abonnements) ; il s'ancre également dans des jeux sociaux de récompense (votes, scores, médailles, gamification), dans une optique d'exposition de soi (comptes non anonymes, ou inversement culture du pseudonyme), ou dans un devoir de réciprocité lorsque l'on a conscience des services que rend la plateforme[8] ;
- viennent enfin les utilisateurs consommateurs, les *free riders* qui profitent des trois premières catégories sans contribuer, hormis en nourrissant un système de consommation de masse qui alimente en notoriété les contributeurs principaux (certains modélisateurs les appellent des « *lemmings* »).

[8] Sur ces aspects voir la revue et le cas empirique présentés par (Guan et al. 2018) – autant d'aspects qui rappellent la façon dont les plateformes contribuent à « enterrer le social sous la simulation du social » suivant l'expression qu'employait Jean Baudrillard (Baudrillard 1982)

Conclusion

Une première partie de cet article s'est centrée sur le jeu de concurrence entre deux logiciels ayant peu ou prou les mêmes fonctions, l'un ayant dix ans d'ancienneté de plus que le second. En l'occurrence R, le plus ancien, et Pandas – Python, le plus récent. La seconde partie s'est penchée sur le déroulé de la croissance de Pandas, le plus récent des deux uniquement (la base de données ne nous permet pas de faire la même chose pour le premier). Nous pouvons résumer nos résultats en cinq points principaux :

- ce projet partait de l'idée d'observer des conversions individuelles, c'est-à-dire des individus changeant d'environnement logiciel durant leur carrière, en partant d'un outil de *data* ancien vers un outil plus moderne, intégré dans un environnement plus généraliste, donc supérieur au premier. Notre principal constat est que les utilisateurs qui font ce genre de conversion sont excessivement rares. Nous y voyons des implications épistémologiques importantes : si ce constat est vrai pour un simple outil de programmation, qu'en est-il pour des éléments de doctrine théorique ou pour des paradigmes plus fondamentaux ? Le renouvellement générationnel semble un meilleur moteur du changement que l'acteur individuel lui-même ;
- un deuxième point illustrant la forte viscosité sociale de la plateforme a été abordé dans la seconde partie. Les individus changeant de rôle dans le réseau sont, eux aussi, très rares. Les individus experts le sont dès le début et le restent. Ceux dépendants du savoir des autres en font autant. Les rôles sont fixes sur la plateforme, et cette recherche n'est pas parvenue à observer de véritables trajectoires de mobilité ascendante. Ce qui nous amène au point suivant ;
- le troisième point illustre une nouvelle fois un résultat contradictoire par rapport aux *a priori* de cette recherche. Alors que nous nous attendions à observer une communauté sans hiérarchie et aux échanges multidirectionnels, c'est-à-dire un réseau classique avec ses cliques, ponts, individus centraux, nous avons trouvé un système dual dans lequel les conseils sont descendants, très majoritairement le fait d'une couche d'experts ne communiquant pas entre eux, mais répondant à une couche de novices en situation de dépendance continue ;
- un quatrième point et dernier point a trait à la production de savoirs sur ces plateformes. Comme on l'a vu avec le rôle central joué par les promoteurs-propriétaires de Pandas, les connaissances qui arrivent sur la plateforme émanent dans un premier temps de l'extérieur, d'organisations classiques (on a vu notamment la centralité des diplômés du MIT). Les communautés SO apparaissent comme des espaces sous influence d'organisations classiques et ne sont pas, dans le cas observé ici, le moteur principal de ce qu'elles produisent.

Ces résultats largement contre-intuitifs montrent l'importance de toute démarche empirique permettant de défaire certains mythes sur le fonctionnement du web. L'exemple de *StackOverflow* nous pousse ainsi vers le constat selon lequel certains réseaux sociaux numériques s'analysent mieux dans une perspective d'écologie des populations que dans une perspective d'analyse de réseaux sociaux. En effet, il est difficile de dégager des mécanismes « réseaux » dans le système hiérarchisé, dual et hétéronome que nous avons décrit dans cette recherche. Alors que les modèles mathématiques abstraits, sans terrain ni data, tels les modèles à effet de seuils discutés ici, occupent une grande partie de l'espace scientifique, voire ses sphères les plus primées, ces résultats viennent

montrer à quel point ils peuvent parfois s'avérer loin de la réalité du fonctionnement des mécanismes sociaux qu'ils cherchent à décrire.

Bibliographie :

Baudrillard, Jean. 1982. *A l'ombre des majorités silencieuses ou la fin du social/L'extase du socialisme*. 1re Collector édition. Paris: 1978.

Chakraborty, Partha, Rifat Shahriyar, Anindya Iqbal, et Gias Uddin. 2021. « How do developers discuss and support new programming languages in technical Q&A site? An empirical study of Go, Swift, and Rust in Stack Overflow ». *Information and Software Technology* 137:106603. doi: 10.1016/j.infsof.2021.106603.

Chaoui, Amal, Sébastien Delarre, Fabien Eloire, Maxime Morge, et Antoine Nongaillard. 2022. « Toward an Agent-Based Model of Community of Practice: Demonstration ». P. 467‑72 in *Advances in Practical Applications of Agents, Multi-Agent Systems, and Complex Systems Simulation. The PAAMS Collection*. Vol. 13616, *Lecture Notes in Computer Science*, édité par F. Dignum, P. Mathieu, J. M. Corchado, et F. De La Prieta. Cham: Springer International Publishing.

Chiang, Yen-Sheng. 2007. « Birds of Moderately Different Feathers: Bandwagon Dynamics and the Threshold Heterogeneity of Network Neighbors ». *Journal of Mathematical Sociology* 31(1):47‑69. doi: 10.1080/00222500601013536.

Delarre, Sébastien, Fabien Eloire, Antoine Nongaillard, et Maxime Morge. 2023. « Modèle explicatif de la sécession des experts dans les communautés de pratiques ». *HAL Id: hal-04164769*.

Fligstein, Neil. 2002. *The Architecture of Markets: An Economic Sociology of Twenty-First Century Capitalist Societies*. 2. print., and 1. paperback print. Princeton, NJ: Princeton Univ. Press.

Fuller, Matthew, Andrew Goffey, Adrian Mackenzie, Richard Mills, et Stuart Sharples. 2016. « Chapter Three. Big Diff, Granularity, Incoherence, and Production in the Github Software Repository ». P. 87‑102 in *Chapter Three. Big Diff, Granularity, Incoherence, and Production in the Github Software Repository*. Amsterdam University Press.

Galeotti, Andrea, et Sanjeev Goyal. 2009. « Influencing the Influencers: A Theory of Strategic Diffusion ». *The RAND Journal of Economics* 40(3):509‑32. doi: 10.1111/j.1756-2171.2009.00075.x.

Galeotti, Andrea, et Sanjeev Goyal. 2010. « The Law of the Few ». *The American Economic Review* 100(4):1468‑92.

Granovetter, Mark. 1978. « Threshold Models of Collective Behavior ». *American Journal of Sociology* 83(6):1420‑43.

Granovetter, Mark, et Roland Soong. 1986. « Threshold models of interpersonal effects in consumer demand ». *Journal of Economic Behavior & Organization* 7(1):83‑99. doi: 10.1016/0167-2681(86)90023-5.

Guan, Tao, Le Wang, Jiahua Jin, et Xiaolong Song. 2018. « Knowledge Contribution Behavior in Online Q&A Communities: An Empirical Investigation ». *Computers in Human Behavior* 81:137‑47. doi: 10.1016/j.chb.2017.12.023.

Henderson, David, et Peter Graham. 2017. « Epistemic Norms and the “Epistemic Game” They Regulate: The Basic Structured Epistemic Costs and Benefits ». *American Philosophical Quarterly* 54(4):367‑82.

Lazega, Emmanuel. 2003. « Rationalité, discipline sociale et structure ». *Revue française de sociologie* 44(2):305‑29. doi: 10.3917/rfs.442.0305.

Macy, Michael W. 1991. « Chains of Cooperation: Threshold Effects in Collective Action ». *American Sociological Review* 56(6):730‑47. doi: 10.2307/2096252.

Macy, Michael W., et Anna Evtushenko. 2020. « Threshold Models of Collective Behavior II: The Predictability Paradox and Spontaneous Instigation ». *Sociological Science* 7:628‑48. doi: 10.15195/v7.a26.

McKinney, Wes. 2011. « Pandas: A Foundational Python Library for Data Analysis and Statistics ».

McKinney, Wes. 2012. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O’Reilly Media, Inc.

Morge, Maxime. 2023. « MoCiCoS / SoDyOnStack · GitLab ». *GitLab*. Consulté 30 janvier 2024 (https://gitlab.univ-lille.fr/mocicos/sodyonstack).

Poussou-Plesse, Marielle, Duplan Denis, Perrin-Joly Constance, et Anne-Marie Guillemard. 2008. « Durer au travail dans les métiers de l’informatique: quelles conditions de possibilité ? Etude sociologique du devenir des cadres informaticiens ». *DARES : Centre d’études des mouvements sociaux*.

Ragouet, Pascal. 2000. « Notoriété Professionnelle Et Organisation Scientifique ». *Cahiers Internationaux de Sociologie* 109:317‑41.

Schelling, Thomas C. 1971. « Dynamic models of segregation† ». *The Journal of Mathematical Sociology* 1(2):143‑86. doi: 10.1080/0022250X.1971.9989794.

Sørensen, Aage B. 2000. « Toward a Sounder Basis for Class Analysis ». *American Journal of Sociology* 105(6):1523‑58. doi: 10.1086/210463.

Sun, Michael Mu, Akash Ghosh, Rajesh Sharma, et Sandeep Kaur Kuttal. 2018. « Birds of a Feather Flock Together? A Study of Developers’ Flocking and Migration Behavior in GitHub and Stack Overflow ».

Übler, Hannah, et Stephan Hartmann. 2016. « Simulating Trends in Artificial Influence Networks ». *Journal of Artificial Societies and Social Simulation* 19(1):2.

White, Harrison C. 2002. *Markets from Networks: Socioeconomic Models of Production*. Princeton University Press.