

Rapport de projet - Data Mining

Timothé BONHOURE - Jocelyn BOURNIER-ARME NI

1 Introduction

L'objectif du projet était de réaliser une analyse sur un jeu de données réel, en utilisant les outils et techniques appropriés aux sujets que nous choisirions d'aborder. Le dataset que nous avons sélectionné est titré : “**Reddit 2023 r/place tile placement data**”. L'événement r/place était une expérience sociale interactive qui a eu lieu sur la plateforme reddit, permettant aux utilisateurs du site de collaborer à la création d'**une grande grille de pixels**. Chaque utilisateur avait la possibilité de **placer un pixel toutes les cinq minutes sur la toile**, résultant en une œuvre collective en constante évolution. Le jeu de données que nous avons choisi d'explorer contient les informations relatives aux placements de pixel tout au long de l'événement. Une ligne du dataset contient pour un pixel posé, la date (au millième près), le nom d'utilisateur (hashé), l'emplacement et la couleur. Nous avons choisi, pour des raisons de performance, de nous concentrer sur un échantillon des données, toutes nos analyses seront ainsi réalisées en utilisant les deux premiers tiers des pixels placés. Au cours de l'événement, la grille a été rallongée plusieurs fois, dans l'échantillon traité, elle a été étendue une fois de chaque côté, sur la verticale. Il a aussi été ajouté de nouvelles couleurs à la palette proposée.

Nous commencerons par une présentation des chiffres et statistiques générales sur le dataset. Ceci nous permettra de dimensionner les données et observer l'évolution de l'activité au cours du temps dans l'échantillon étudié. Ensuite, nous passerons à une partie concernant la recherche de différentes communautés qui ont pu se rassembler pour représenter des motifs particuliers sur la grille.

2 Statistiques et présentation générale du dataset

Sur l'échantillon traité, nous avons un total de presque 50 millions de pixels placés, dont 3 millions par les modérateurs pour retirer des motifs considérés comme inappropriés.

Une représentation graphique de la répartition spatiale de l'ensemble de l'activité sur le r/place, avec un delta d'importance par pixel placé proportionnellement plus grand pour les zones apparues plus tard, est visible dans la figure 1 ci-dessous. On y constate des zones où différentes communautés ont dû rentrer en conflit pour faire apparaître et maintenir leur motif sur la grille.

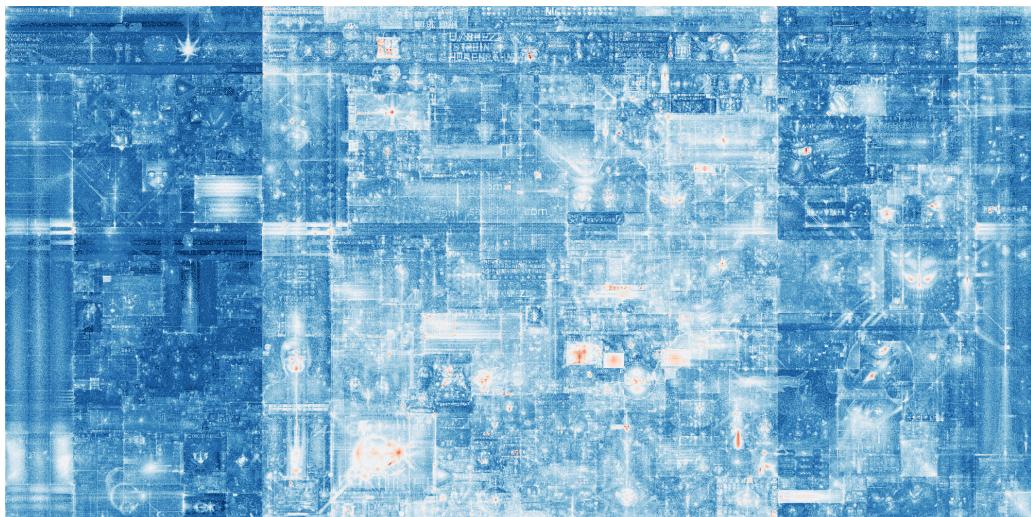


FIGURE 1 – Activité totale, de bleu (inactif) à rouge (très actif)

Nous supposions et avons pu vérifier que le nombre de pixels placés par utilisateur, modélisant leur niveau d'investissement dans l'événement, devrait suivre une distribution en loi de puissance. Cette hypothèse s'est fortement confirmé après agrégation du nombre de pixels posés pour chaque utilisateur actif sur l'échantillon traité. La loi de distribution que nous avons obtenue est présentée dans la figure 2.

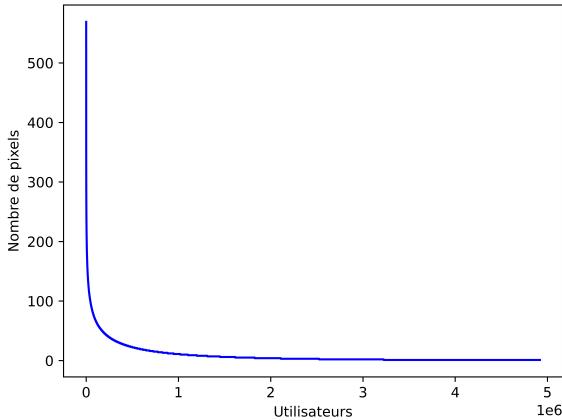


FIGURE 2 – Fréquence de placement de pixels par utilisateur

Nous avons aussi cherché à observer l'évolution de l'activité au cours de l'événement, tout en surveillant la répartition des différentes couleurs placées pour chacun des points, pour ce faire, nous avons réalisé plusieurs vues, la première, en figure 3, donne en chaque point un suivi du nombre de pixels posés sur les 5 dernières minutes, avec une moyenne lissé sur les 10 dernières minutes. Chaque aire sous la courbe d'une couleur représente le pourcentage de points placés de cette couleur à un instant t . On peut constater des creux qui se sont répétés sur les deux jours, représentant sûrement les horaires d'inactivités les plus probables pour les pays les plus impliqués, sachant que l'événement a démarrer à 13H UTC.

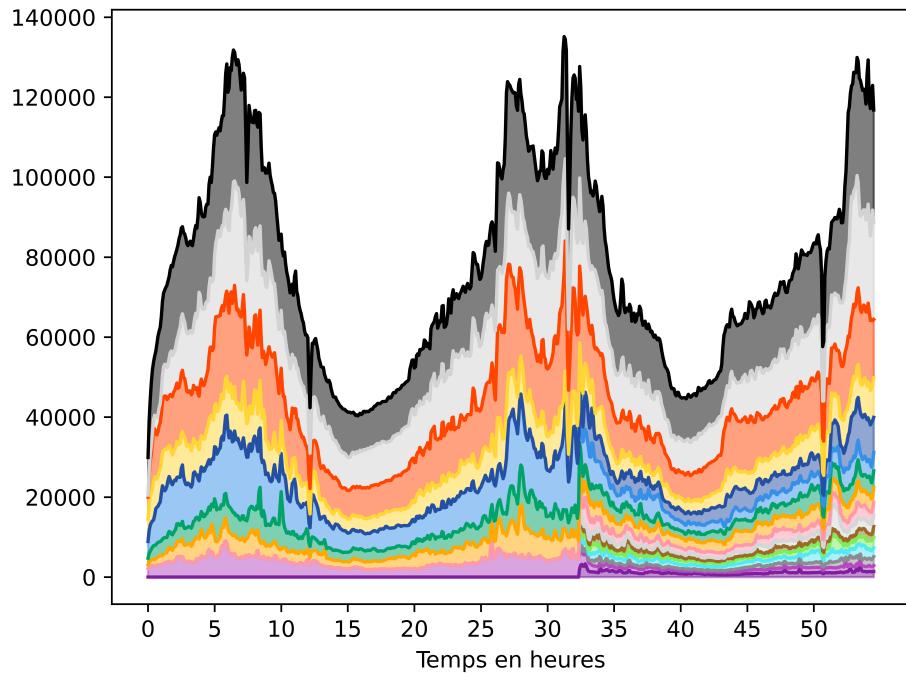


FIGURE 3 – Moyenne glissante du nombre de pixels placés sur les 5 dernières minutes, par couleur

La seconde (figure 4), plus localisée sur le début de l'événement, permet de constater que les utilisateurs étaient plutôt synchronisés sur le placement des pixels au début, puis bien sûr une désynchronisation est vite survenue.

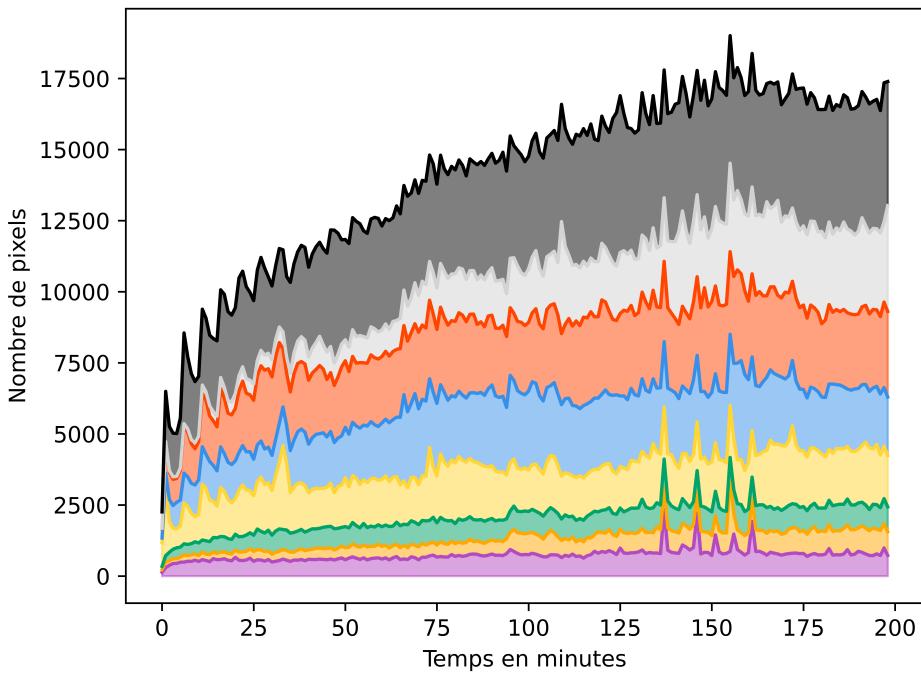


FIGURE 4 – Début de la fréquence de placement par couleur

Enfin, nous avons constaté en travaillant sur les données, une anomalie survenue au bout d'environ x heures, nous n'avons pas pu définir avec certitude la raison de cette anomalie, mais nous avons quelques hypothèses. Cette anomalie se présente par une chute dans l'activité des utilisateurs que l'on observe dans le zoom présenté en figure 5. Durant l'événement, il y a eu plusieurs soucis de serveurs, mais non globaux en général, il est possible qu'un problème technique soit survenu pour tout le monde sur un intervalle de temps court. Ou il est possible que le souci vienne de l'enregistrement des données, peut-être que certains points placés ont été mal enregistrés.

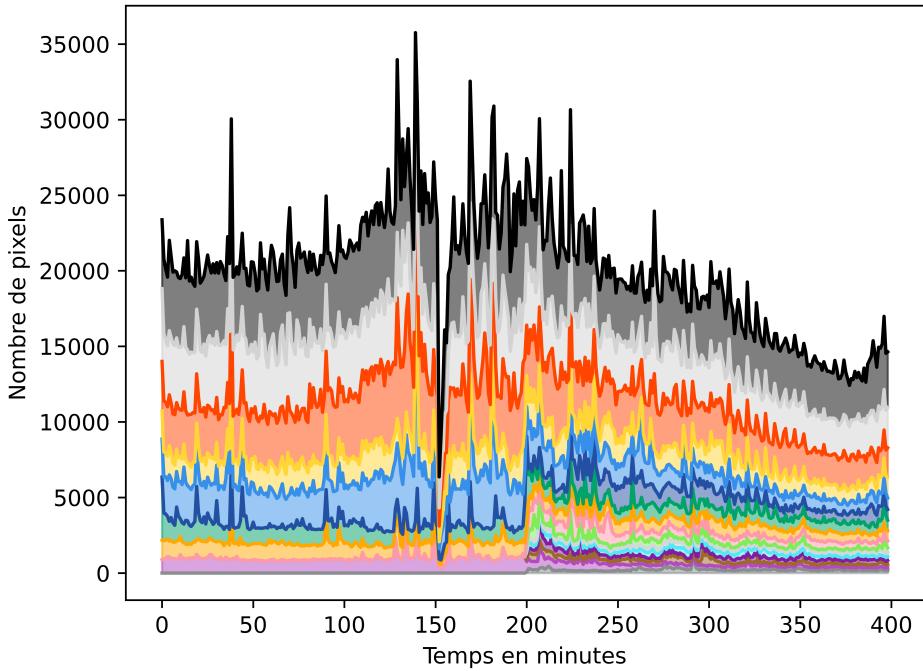


FIGURE 5 – Vue rapprochée de l'anomalie de la fréquence de placement

3 Recherche de communautés

Nous avons chercher à détecter des communautés d'utilisateur. Pour Commencer, nous avons tout d'abord détecter des utilisateurs étant en opposition. Deux utilisateurs ayant placé l'un après l'autre sur le même pixel considéré en opposition. Nous avons fait ce choix car lorsqu'un utilisateur mets un pixel juste après un autre c'est qu'il souhaite changé la couleur que le précédent utilisateur avait mis. Nous avons de plus pondéré les liens entre les utilisateurs en fonction du nombre de fois que cela était arriver

Pour détecter des communauté à partir de ce graphe nous avons utilisé une approche de type l'ennemie de mon ennemis est mon ami. Pour calculer le lien entre deux noeud x,y nous avons regarder l'ensemble $CE(x,y)$ des ennemis commun aux deux noeud mais aussi si eux même étaient ennemis.

4 Conclusion