

Analiza sieci społecznościowej

Piotr Wojdas - inżynieria systemów, 3 rok

Wybrana sieć - Najpopularniejsze hasztagi na X

Będziemy rozpatrywać sieć najpopularniejszych hashtagów, które miały swój "prime" w określonym czasie. Zbiór danych pochodzi ze strony kaggle i zawiera informacje o tym jaki tag miał swój szczyt popularności w określonym czasie

Podstawowe informacje o sieci:

- Węzły - unikalne hashtagi
- Krawędzie - Połączenie między hashtagami, jeśli miały one swój szczyt popularności w podobnym okresie (2 tygodnie)
- Będziemy analizować, co można odczytać na podstawie występujących hashtagów na platformie X oraz ich popularności w danym okresie czasowym

Statystyki:

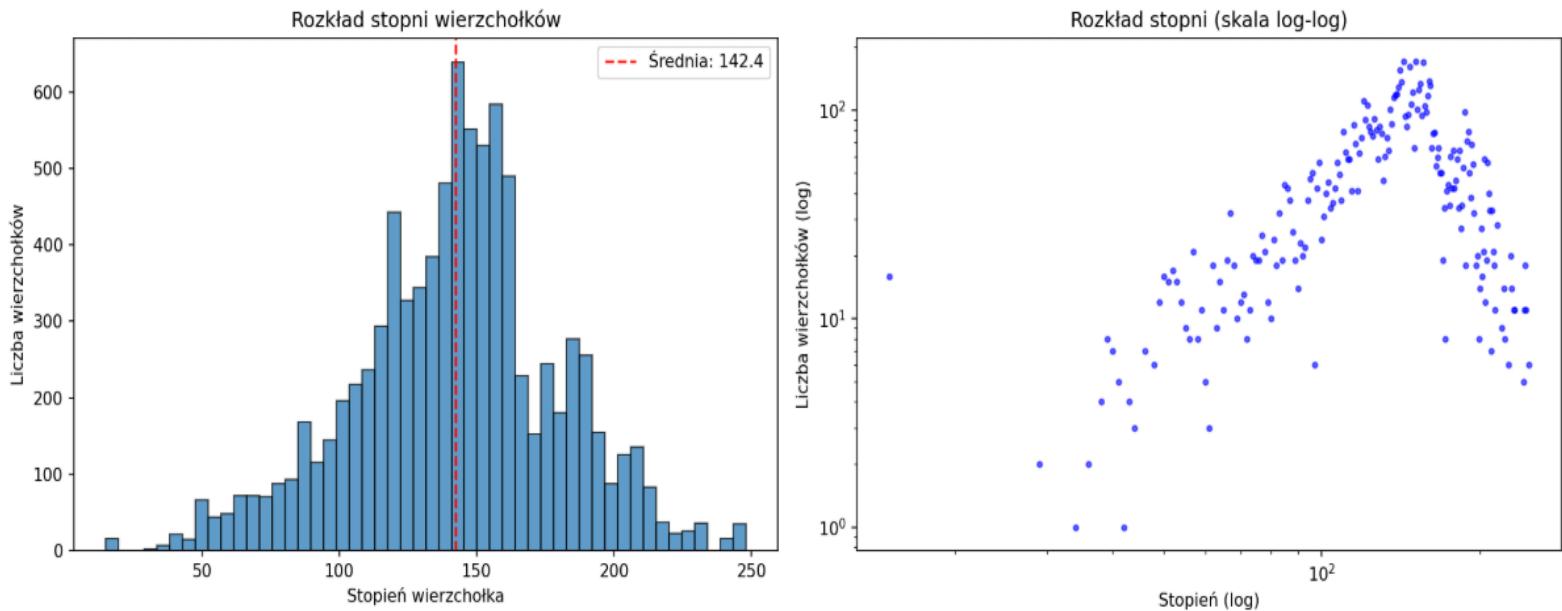
- Liczba wierzchołków - 8861
- Liczba krawędzi - 630713
- Gęstość grafu - 0.016 (nieco ponad 1%)

Możemy zauważyć, że sieć ta jest bardzo rzadka. Dzieje się tak głównie dlatego iż szczyty popularności hashtagów rozkładają się w czasie, nie może za wiele rzeczy być na szczytach, ludzie mają tendencję do skupiania się na jednej rzeczy, odstawiając inne w kąt, stąd na przestrzeni 6 lat, ponad 7 tysięcy wątków miało swój moment sławy, gdzie tylko niewielki procent występował w tym samym czasie. Warto wspomnieć też, że graf ten **JEST SPÓJNY** więc na przestrzeni tych wszystkich lat nie było 2 tygodni, w których nie byłoby o czym dyskutować (pokazuje to też że nie było sytuacji gdzie twitter był nieczynny na dłużej niż ten okres)

- Średnica - 119

Średnica pokazuje nam tylko tyle, że graf ten jest **łańcuchem czasowym** i udowadnia fakt, że sieć pokazuje, jakie tematy były popularne w tym samym czasie.

Rozkład stopni wierzchołków:

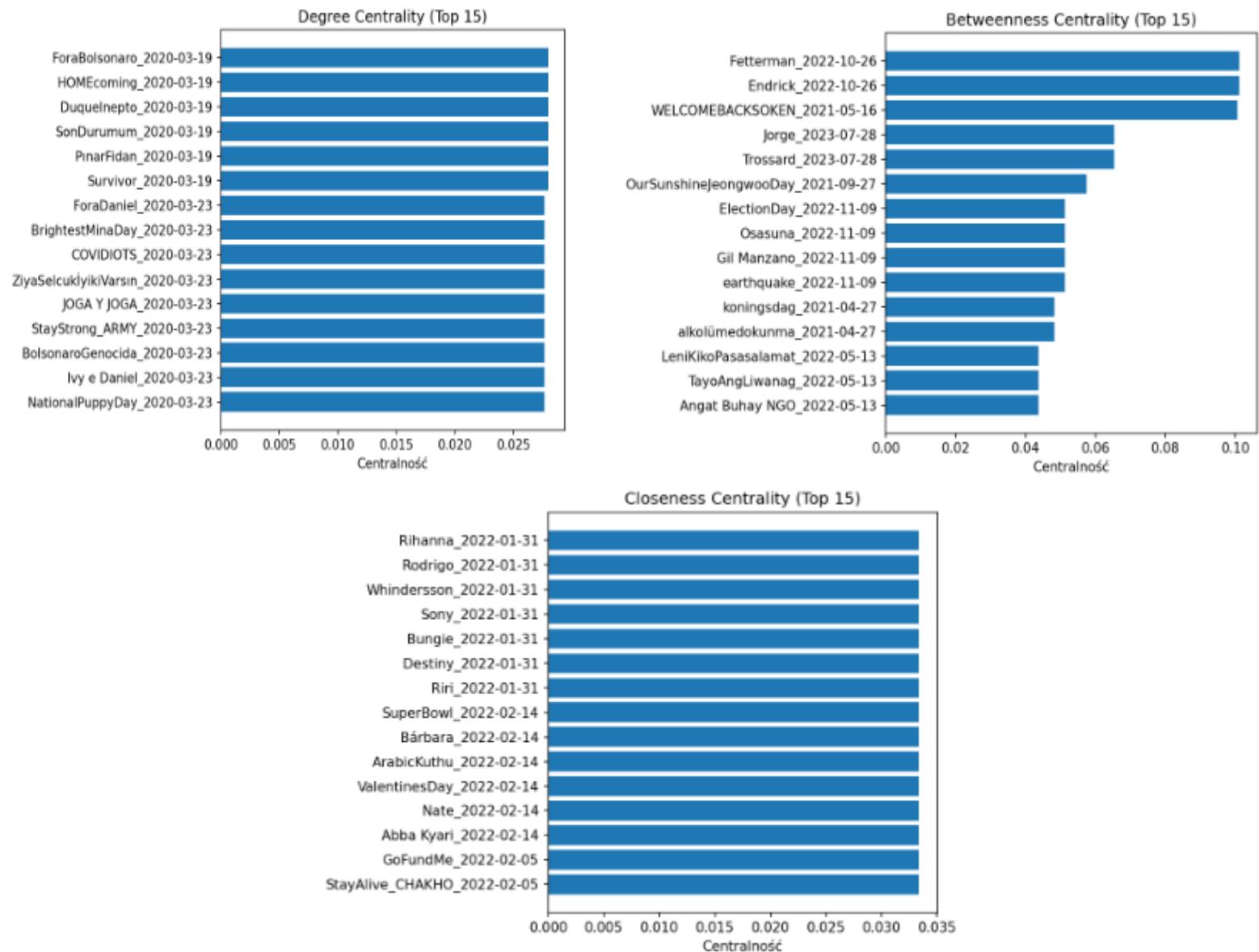


Z rozkładu stopni widzimy, że ma on rozkład przypominający normalny, gdzie zdecydowana większość wierzchołków ma w okolicach 142 krawędzi. Mówiąc nam to mniej więcej tyle, że na przestrzeni lat liczba popularnych wątków nie zmienia się drastycznie, jest stabilna. Wartości poniżej średniej są spowodowane głównie przez zamknięte okno czasowe, gdzie wierzchołki na granicach (2020 i 2025 rok) naturalnie tych połączeń będą mieć mniej, z kolei wartości powyżej średniej to okresy, gdzie działa się wyjątkowo dużo, co potwierdza **top 10 tagów wg stopnia**:

1. ForaBolsonaro_2020-03-19: 248
2. HOMEcoming_2020-03-19: 248
3. DuquelNepto_2020-03-19: 248
4. SonDurumum_2020-03-19: 248
5. PınarFidan_2020-03-19: 248
6. Survivor_2020-03-19: 248
7. ForaDaniel_2020-03-23: 245
8. BrightestMinaDay_2020-03-23: 245
9. COVIDIOTS_2020-03-23: 245
10. ZiyaSelçukluVarsın_2020-03-23: 245

Wśród nich mamy chociażby tagi takie jak COVIDIOTS oraz SURVIVOR, które nawiązują do okresu pandemii, która z kolei z pewnością generowała duże poruszenie wśród społeczności (przymusowe siedzenie w domach zwiększyło aktywność w social mediach co pewnie też w jakiś sposób wpłynęło na takie natężenie wątków w tym okresie).

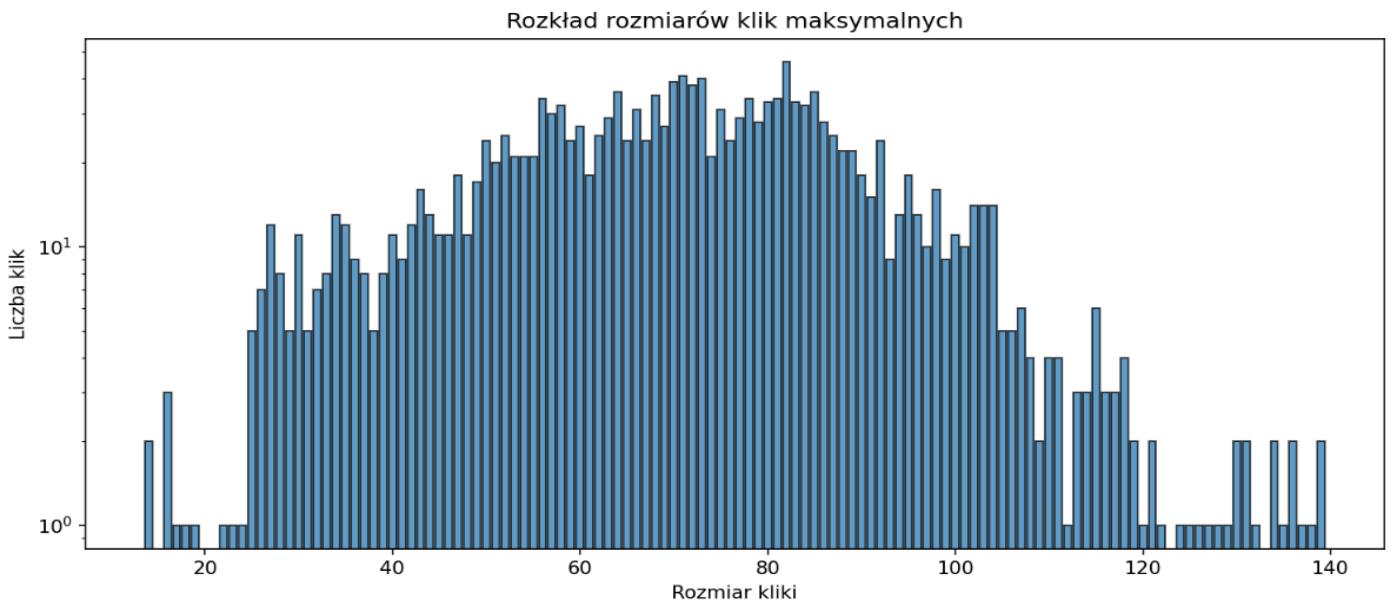
Miary centralności:



W naszym przypadku miary centralności zmieniają lekko swoje znaczenie.

- Centralność stopnia: Już omówiona, najczęściej trendów było w trakcie pandemii
- Centralność pośrednictwa: Pokazuje nam hasztagi, które były popularne, gdy “mało się działało”, były pomiędzy intensywnymi okresami
- Centralność bliskości: Tutaj wszystkie węzły są w okolicach początku lutego 2022 co jest środkiem jeśli chodzi o okres czasowy zbierania danych (2020-2025) więc te węzły mają najbliższej do początku i końca tego okresu

Kliki:

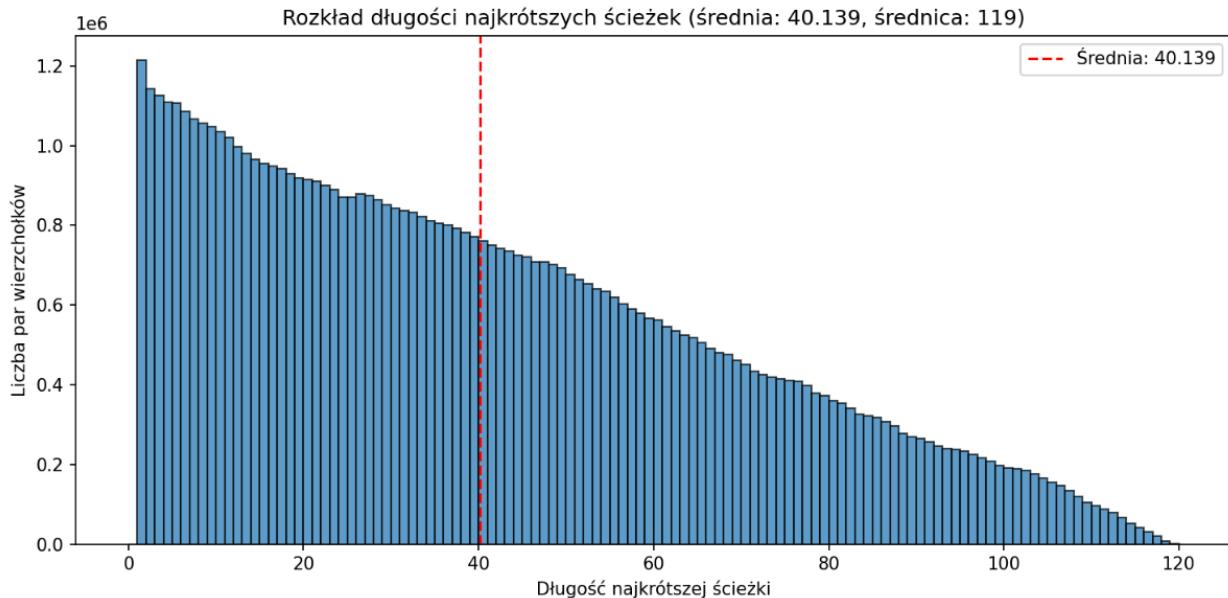


Sieć ta posiada dużą liczbę klik co jest normalne, gdyż kliką jest każde okno czasowe o szerokości 2 tygodni. W środku takiego okna wszystkie wierzchołki są ze sobą połączone więc tworzą klikę.

Największa klika to 139 hashtagów, występujących naraz, znowu jest to odniesienie do pandemii, podczas której miało to miejsce, mały rozmiar klik ponownie też przypada na tygodnie skrajne.

Średnio na przestrzeni 2 tygodni przypada 70 tematów, które zyskują popularność co nie budzi większych wątpliwości przy zestawieniu z innymi statystykami

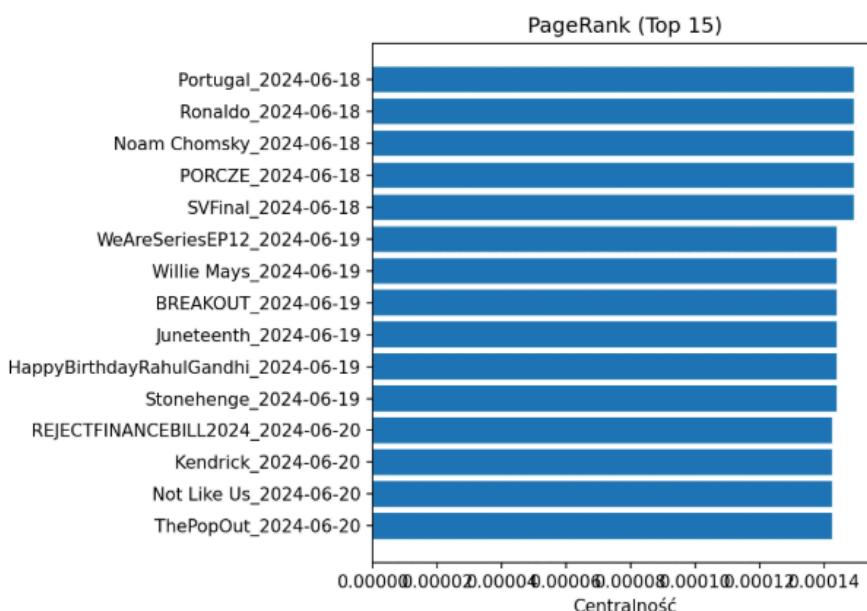
Rozkład długości najkrótszych ścieżek



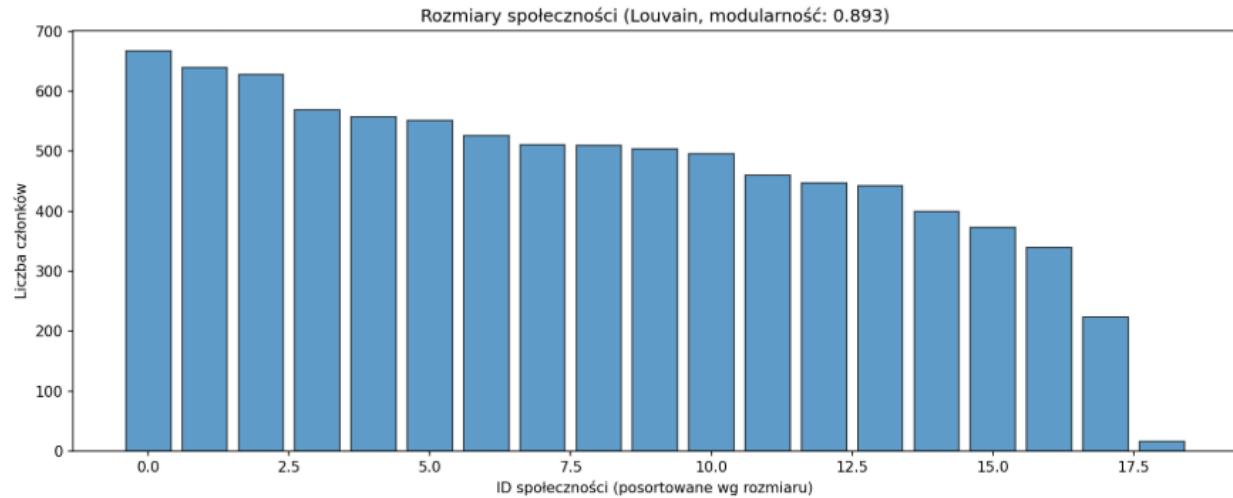
Ten rozkład nie pokazuje akurat już nic nowego, duża liczba klik łączy się z faktem, że dużo wierzchołków łączy się ze sobą przy użyciu niewielu krawędzi i własność ta spada liniowo wraz z odlegością czasową wierzchołków (w końcu czas jest liniowy)

Luki w danych

Na przestrzeni analizy sieci zauważałem fakt, iż występują gdzieś luki w danych. Niektóre miesiące bądź tygodnie posiadają bardzo niską liczbę wierzchołków (co dolicza się do wykresów przy małych liczbach klik/stopni itp). Z tego chociażby powodu, PageRank wskazuje na eventy, które dzieją się w czerwcu 2024, ponieważ po tym okresie wystąpiła luka danych, gdzie w lipcu tego samego roku było bardzo mało wierzchołków, stąd ta miara centralności zatrzymała się i nie doszła do końca sieci.



Społeczności



Wykorzystałem algorytm louvain, który znajdował w tej sieci 19 społeczności

Top 10 największych społeczności:

1. Społeczność 16: 668 członków
Przykłady: ['BBMAsTopSocial_2021-05-11', 'PermissiontoDance_2021-07-09', 'Tokyo2020_2021-07-24', 'popcat_2021-08-15', 'Cuba_2021-07-12']
2. Społeczność 17: 640 członków
Przykłady: ['Kobe_2020-01-27', 'KickIt1stwin_2020-03-27', 'Bernie_2020-04-08', 'Marcela_2020-04-06', 'GRAMMYs_2020-01-27']
3. Społeczność 12: 628 członków
Przykłady: ['GRAMMYs_2022-04-04', 'Eurovision_2022-05-14', 'Elon Musk_2022-04-26', 'Oscars_2022-03-28', "Father's Day_2022-06-20"]
4. Społeczność 5: 569 członków
Przykłady: ['Valentine's Day_2024-02-14', 'Happy New Year_2024-01-01', 'Oscars_2024-03-11', 'Epstein_2024-01-04', 'Chiefs_2024-02-12']
5. Społeczność 6: 557 członków
Przykłady: ['Trump_2023-04-05', 'deprem_2023-02-06', 'MetGala_2023-05-02', 'Oscars_2023-03-13', 'Ahbap_2023-02-10']
6. Społeczność 18: 551 członków
Przykłady: ['Trump_2020-09-30', 'Halloween_2020-10-31', 'COVID_2020-10-02', 'Biden_2020-11-04', 'RRBExamDates_2020-09-05']
7. Społeczność 14: 526 członków
Przykłady: ['BTS BTS_2021-10-02', 'BTSxAMAs_2021-11-22', 'Thanksgiving_2021-11-26', 'Spotify_2021-12-02', 'WhatsApp_2021-10-05']
8. Społeczność 4: 511 członków
Przykłady: ['MetGala_2024-05-07', 'Trump_2024-05-30', 'Israel_2024-04-14', 'Biden_2024-06-28', "Mother's Day_2024-05-13"]
9. Społeczność 13: 510 członków
Przykłady: ['Ukraine_2022-02-24', 'PTD_ON_STAGE_SEOUL_2022-03-10', 'Putin_2022-02-24', 'BBB22_2022-02-15', 'Happy New Year_2022-01-01']
10. Społeczność 15: 504 członków
Przykłady: ['StanWorld_2021-03-31', 'Trump_2021-01-07', 'GRAMMYs_2021-03-15', 'America_2021-01-07', 'Capitol_2021-01-07']

Społeczności te nie zostały jednak podzielone na grupy tematyczne (na przykład na politykę, sport itp.) a na czasowe, widać tutaj że w jednej społeczności hashtagi różnią się od siebie tematyką, ale pochodzą z bliskiego sobie okresu. Pokazuje to, że o ile w szerokim rozumieniu, liczba trendów jest stała, to w praktyce występują miesiące bardziej intensywne i te mniej, choć intensywnych jest zdecydowanie więcej.