

DYNAMIKA TRIAD W SERWISIE INSTAGRAM

TRIAD DYNAMICS ON INSTAGRAM

Streszczenie: Ewidencja triad jest uznanym sposobem opisu sieci złożonych. W pracy przedstawiono analizę dynamiki transformacji tzw. triady zakazanej w sieci społecznej użytkowników serwisu Instagram. Uzyskane wyniki wskazują, że szansa domknięcia takich triad do pełnych trójkątów maleje z czasem, i że ogólnie jest ona tym większa im wyższy jest stopień węzłów wchodzących w jej skład. Ponadto, triady mało powiązane z resztą sieci pozostają mniej wrażliwe na trendy i anomalie o zasięgu globalnym.

Abstract: Triad census remains a recognized approach to describe complex network structures. Dynamics of transition of a so-called forbidden triad into a full triangle, is analyzed and presented. The dataset used is a newly downloaded section of Polish Instagram network. The results indicate that the triad transition probability decreases with time, while it is generally higher for triads that remain better connected with the rest of the network. On the contrary, transition of triads with small-degree nodes remains unaffected by network-wide trends and anomalies.

Słowa kluczowe: analiza sieci społecznych, ewidencja triad, zakazana triada.

Keywords: social network analysis, triad census, forbidden triad.

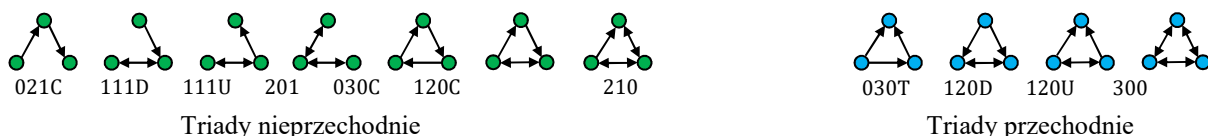
1. WSTĘP

Systemy o strukturze sieci złożonych, pierwotnie obecne naturalnie w procesach fizycznych, chemicznych i biologicznych, w miarę rozwoju społeczeństw powstawały również w wyniku planowej lub spontanicznej aktywności człowieka – technologicznej bądź społecznej. Obecnie, dzięki nowoczesnym technologiom, rozmaite sieci społeczne opłoty jednostkę, wpływając bardziej niż kiedykolwiek na jej decyzje.

Siłę i charakter wzajemnych oddziaływań w takich sieciach próbuje się przedstawiać w sposób syntetyczny, w odniesieniu do systemu jako całości, poprzez zestaw parametrów charakterystycznych: średnicy, współczynnika gronowania, współczynnika rozkładu potęgowego stopnia węzłów, asortatywności i paru innych. Natomiast w odniesieniu do poszczególnych węzłów operujemy różnymi współczynnikami centralności, pojęciami mostów i przegubów etc. Wspomniane pojęcia dotyczą i obliczane są w odniesieniu do konkretnej, niezmiennej struktury sieciowej – ale tak naprawdę próbują one opisać zjawiska dynamiczne w sieci: przepływy pomiędzy węzłami, ich polaryzację, czy wreszcie zmianę struktury samej sieci. Na przykład, wartość współczynnika rozkładu potęgowego jest ściśle powiązana z rozkładem prawdopodobieństwa w algorytmie generacji sieci metodą losowego, preferencyjnego wiązania się nowych węzłów z istniejącymi [1], a centralność wg wektorów własnych – z prawdopodobieństwem przebywania w węzle procesu losowo błądzącego w sieci.

Wygodną techniką analizy sieci, zaproponowaną w [9], jest analiza triad, czyli wszystkich podgrafów trójwierzchołkowych. Częstość występowania określonych typów triad lub klas triad ma ścisły związek ze wspomnianymi już globalnymi parametrami sieci. Na przykład, częstość występowania triad przechodnich (por. Rys. 1) służy do obliczenia współczynnika gronowania.

Ewidencja triad w grafie (*triad census*), czyli wyznaczenie procentowego udziału wybranych typów triad, to uznany od wielu lat sposób opisu aktualnego stanu grafu. Choć już stosunkowo dawno temu [5] zaproponowano i uzasadniono potrzebę analizowania podgrafów bardziej złożonych niż triady, ewidencja triad wciąż pozostaje preferowaną techniką syntetycznego opisywania sieci. Naukowcy z różnych dziedzin uznają udział

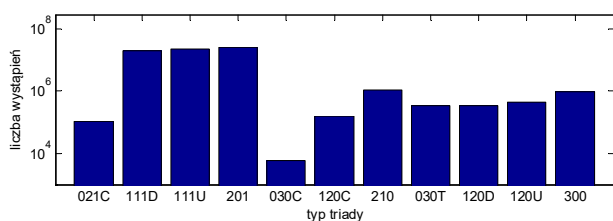


Rys. 1. Sposób oznaczania triad zaproponowany w [9] (na rysunku przedstawiono jedynie triady spójne). Oznaczenie typu triady składa się z liczby relacji symetrycznych, liczby relacji asymetrycznych, liczby relacji nieistniejących oraz, opcjonalnie, kodu literowego ujednoznaczniającego strukturę triady (Up, Down, Cyclic). Triada jest przechodnia, jeśli spełnione są zależności „A wskazuje na B”, „B wskazuje na C” oraz „A wskazuje na C”.

triad określonego typu za wyróżnik relacji społecznych w sieciach. Na przykład, w odniesieniu do interakcji zwierząt stadnych [3], dominacja triad typu 300 jest ewidentna w przypadku iskania oraz walk szympansov, nieantagonistycznych relacji gerez (*Colobus*) i dominacji u pawianov; natomiast ponadprzeciętna obecność triad 030T, 030C i 021C charakteryzuje sieci makaków i koczokodanov (a konkretnie, walki pomiędzy rocznymi osobnikami). Ta sama autorka, wskazując typy triad nosące najwięcej informacji o sieci [4], cytuje przykłady zastosowania triad jako elementu analizy tendencji samobójczych [2]. Ewidencja triad okazuje się narzędziem przydatnym również w analizie sieci współpracy ludzi lub organizacji [11], umożliwiając wychwycenie struktur zhierarchizowanych (triada 201) czy też anomalii w przepływie informacji (triada 030C). Skład procentowy triad jest niezmienny i wyróżnia sieci reprezentujące różne klasy zjawisk: sąsiedztwo słov w językach, powiązania społeczne, powiązania biologiczne [12], przy czym największe znaczenie ma proporcja triad przechodnich wobec nieprzechodnich oraz ilość triad zakazanych w sieci.

Podstawowa i najbardziej uznana reguła dotycząca dynamiki sieci postrzeganej jako zbiór triad, została sformułowana w [7]: w sieci nieskierowanej, gdzie krawędzie reprezentują faktycznie mocne powiązania między węzłami, triady „otwarte” tj. z połączonymi jedynie dwoma węzłami są zjawiskiem przejściowym – takie struktury skazane są albo na zanik jednego z istniejących połączeń, albo na powstanie brakującego połączenia. Dlatego triada tego typu jest określana jako „zakazana”. W terminologii z Rys. 1 taką zakazaną jest triada typu 201, a triadą domkniętą – 300. (Uznajemy tutaj, że postulowane mocne powiązania to powiązania w szczególności odwzajemnione – stąd wybór ww. triad – choć w rzeczywistych sieciach relacje interpersonalne są zdecydowanie asymetryczne.)

Pierwotną motywacją domykania się triad zakazanych jest wzajemne zainteresowanie sobą niepowiązanych dotychczas osobników. Jeśli nie dochodzi ono do skutku, osobnik pośredniczący wykazuje skłonność do przerwania jednego z własnych powiązań z powodu dyskomfortu, jaki odczuwa, pielęgnując relacje z dwoma uparcie niewspółpracującymi swoimi znajomymi. Jak już wspomniano, uzasadnienie takie jest przekonujące w relacjach autentycznie symetrycznych – kontrprzykładem może być relacja agenta nieruchomości i jego klientów, w której pozorna symetria komunikacyjna maskuje faktyczną asymetrię informacyjną. W dalszych rozważa-



Rys. 2. Ewidencja wybranych triad w sieci stworzonej z danych pobranych z serwisu Instagram

niach będziemy zakładać, że struktura grafu odpowiada autentycznym relacjom międzyludzkim, i że w szczególności połączenie dwukierunkowe pomiędzy węzłami oznacza, że żaden z nich nie ma pozycji dominującej.

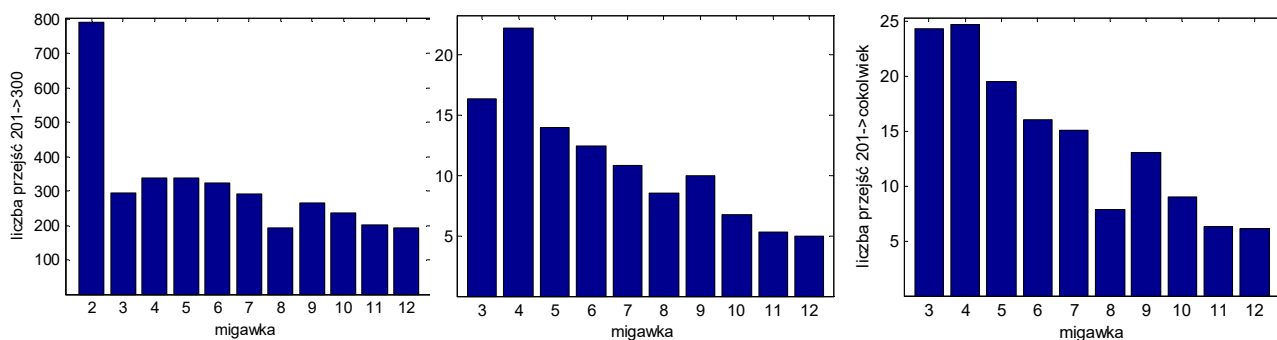
Celem podjętych badań była weryfikacja hipotezy o samoistnym domykaniu się triad zakazanych, na zupełnie nowym zestawie danych testowych. Podstawowe zagadnienie badawcze brzmi więc: co i w jakim stopniu determinuje przeobrażenie się triady typu 201 w triadę typu 300? Znając odpowiedź, można w analogiczny sposób skonstruować model dynamiki sieci w odniesieniu do wszystkich typov triad, co pozwoli prognozować rozwój sieci społecznej, a w szczególności – rozprzestrzenianie się idei, chorovb, powstawanie grup etc.

2. ANALIZOWANE DANE

W badaniach wykorzystano dane o powiązaniach użytkowników serwisu społecznościowego Instagram, pozyskane i wstępnie przeanalizowane w pracy magisterskiej [10]. W pracy postawiono zadanie pozyskania danych z zupełnie nowego źródła, stosunkowo młodego i o wysokiej dynamice. Instagram spełniał wszystkie te cechy: w skali światowej jest jeszcze serwisem dość młodym [6], wciąż ustępując znacznie Facebookowi, YouTube, Google+ i Twitterowi; plasując się mniej więcej na równi z LinkedIn oraz Pinterest. W Polsce miał on w czasie zbierania danych około 3 mln użytkowników [13], z czego większość to kobiety i nastolatki, skupione na przekazie wizualnym. Pomimo wykupienia serwisu przez Facebooka, nowy właściciel nie przeprowadził połączenia kont użytkowników; zatem proces ewolucji sieci społecznej w serwisie Instagram pozostał spontaniczny, niezakłócony.

W pracy [10] pobrano 12 migawek wycinka sieci połączeń pomiędzy użytkownikami, w okresie od końca listopada 2015 r. do końca marca 2016 r. Granice obserwowanego wycinka zostały określone przez własny algorytm autora pracy; rozmiar tego wycinka jest ograniczony liczbą bezpłatnych zapytań do API Instagrama tak, aby pobieranie pojedynczej migawki nie trwało dłużej niż kilka dni. Prace doprowadziły do zgromadzenia informacji o 82 tys. użytkowników (zlokalizowanych głównie w Warszawie), połączonych ponad 2 mln krawędzi, tj. relacją śledzenia wpisov jednej osoby przez inną. Na Rys. 2 przedstawiono procentowy udział wybranych typov triad w obserwowanej sieci, w skali logarytmicznej. Uderza przewaga triad nieprzechodnich, wbrew hipotezie o przejściowości triady zakazanej. Co zatem determinuje trwanie lub domykanie się triad zakazanych?

Posiadany zbiór danych nie zawiera informacji o treściach zamieszczanych przez użytkowników w serwisie Instagram, których analiza mogłaby przyczynić się do wyjaśnienia procesu domykania się triady zakazanej. Dysponujemy jednak pełną informacją o aktualnej strukturze sieci, zatem można zbadać zależność intensywności zjawiska domykania triad od podstawowych parametrav sieciowych poszczególnych węzłov w triadzie, tj. stopnia wejściowego i wyjściowego. Dodatkowo, dysponujemy informacją o liczbie triad



Rys. 3. Z lewej: ogólna intensywność domknięć 201 → 300, w kolejnych migawkach [triada/dzień]. Pośrodku: intensywność domknięć do typu 300 tylko triad 201 nowopowstałych pomiędzy migawką 1. a 2. Z prawej: intensywność dowolnych rodzajów domknięć triad 201 nowopowstałych pomiędzy migawką 1. a 2.

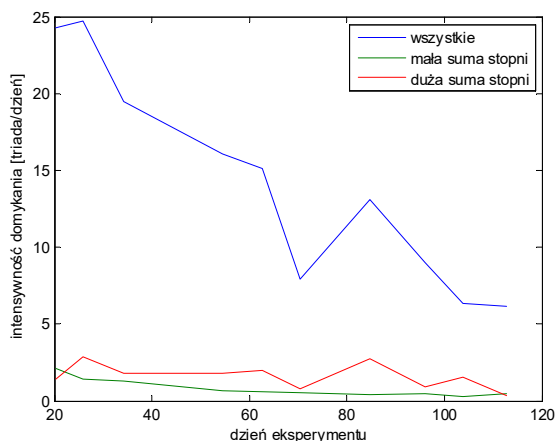
zakazanych, w których uczestniczy wybrana para węzłów. Wpływ tychże parametrów na dynamikę triady zakazanej został zbadany eksperymentalnie.

3. REZULTATY

Dane pobrane z serwisu Instagram opisują liczbę zaobserwowanych nowych połączeń w sieci, w kolejnych sesjach skanowania serwisu. Zatem kolejne migawki wybranego fragmentu sieci składają się z dotychczas zaobserwowanej struktury, plus informacje o nowych połączeniach, pochodzące z ostatniej sesji skanowania. W tak zrekonstruowanej sieci niemożliwe jest wykrycie zaniku odnotowanej już raz relacji międzywęzłowej, dlatego zaobserwowane triady mogą się tylko wzbogacać o nowe połączenia, aż do finalnego domknięcia w typie 300. Dane zgromadzono w relacyjnej bazie SQLite.

3.1. Intensywność domyknięcia w funkcji czasu

Analizę intensywności domyknięcia triad w okresie pomiędzy kolejnymi migawkami przedstawiono na Rys. 3. Jeśli analiza dotyczy wszystkich początkowo zastanych w sieci triad, wówczas okazuje się, że zasadniczo tempo domyknięcia się zależy w niewielkim stopniu od czasu, z wyjątkiem pierwszego okresu. Wynika to jed-



Rys. 4. Intensywność domyknięcia się triad 201, w rozbiciu na triady o małej i dużej sumie stopni wierzchołków.

nak z faktu, że zarejestrowane początkowo triady znajdowały się w różnych stadiach swojej ewolucji, tzn. istniały zróżnicowany okres czasu i dlatego domykały się równomiernie (Rys. 3., lewy). Jeśli natomiast przyglądamy się intensywności zanikania tylko *nowych* triad, tj. powstałych pomiędzy pierwszą a drugą migawką, wówczas widać wyraźnie, że intensywność domyknięcia triad do typu 300 zanika z upływem czasu. Oczywiście, możliwa jest dalsza, dokładniejsza analiza tej dynamiki, z podziałem na typy triad, z których powstały nowe odnotowane triady typu 201.

Pozorna anomalia, polegająca na większej intensywności domyknięcia triad w okresie 4 niż w 3 ma łatwe wytłumaczenie: stadium przejściowym procesu domyknięcia 201 → 300 jest triada 210. Jest ona konfiguracją przejściową tym bardziej, że Instagram zachęca użytkowników, by odwzajemniali relację bycia śledzonym. Uwzględnienie w analizie przejścia triady 201 do dowolnej innej formy (tj. 210 lub 300 – Rys. 3., prawy) daje wykres już z mniejszą nieregularnością na początku. Widzimy więc, że czas życia triady 201 może być modelowany za pomocą rozkładu zbliżonego do trójkątnego, wykładniczego lub beta.

3.2. Intensywność domyknięcia a stopnie wierzchołków triady

W obliczu zaobserwowanego zaniku intensywności domyknięcia się triad warto sprawdzić, czy jest on uniwersalny, czy raczej dotyczy triad 201 o określonych właściwościach. Przeprowadzono zatem ponowną analizę domyknięcia, uwzględniając osobno triady o łącznym małym stopniu wierzchołków i osobno – o dużym łącznym stopniu wierzchołków. Kryteria klasyfikacji związane były z 25. i 75. percentylem w zaobserwowanym rozkładzie stopni wierzchołków triad (traktowanych bez wyróżnienia wierzchołka centralnego triady).

Wyniki analizy przedstawiono na Rys. 4. Widać wyraźnie, że systematyczny zanik intensywności domyknięcia jest właściwy triadom o niewielkim stopniu wierzchołków. Można zaryzykować stwierdzenie, że w tej grupie triad obserwujemy naturalną tendencję do samistnego domyknięcia się triady zakazanej, a nawet wyznaczyć stałą czasową, czy też „czas półtrwania” triad typu 201. Natomiast triady mocniej usieciowione wykazują większe fluktuacje i mniejszą zależność intensywności domyknięcia od upływu czasu. Fluktuacje te wynikają

zapewne z jakichś nierejestrowanych czynników zewnętrznych, które spowodowały ogólnie zmniejszoną aktywność użytkowników w migawce 8 (30 stycznia 2015 r.), a zwiększoną w migawce 9. Można tę anomalię tłumaczyć przejściem z okresu przed feriami szkolnymi na Mazowszu i sesją egzaminacyjną na wyższych uczelniach (migawka 8) do okresu ferii i sesji (migawka 9).

3.3. Intensywność domykania a uczestniczenie w wielu triadach typu 201

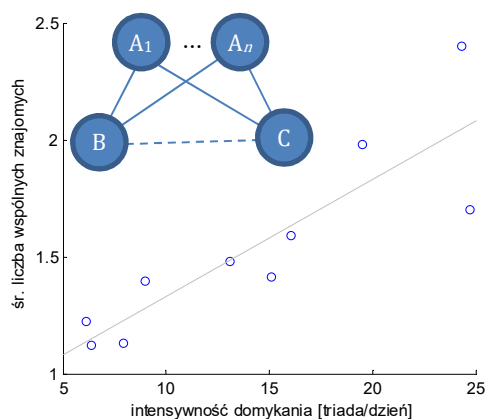
O ile proces domykania się triad o węzłach słabo usieciowionych odbywa się w miarę niezależnie od czynników zewnętrznych (por. Rys. 4, wykres zielony), należy zbadać, co jeszcze determinuje ogólnie obserwowaną, wyższą intensywność domykania dla triad młodych niż starszych. Naturalnym czynnikiem intensyfikującym prawdopodobieństwo domknięcia połączenia B-C w triadzie A-B-C powinien być fakt, że B i C należą jednocześnie do większej liczby triad – por. grafika osadzona w Rys. 5. W Tab. 1 rozbito domykające się w kolejnych migawkach triady według liczby n wspólnych znajomych węzłów B i C. Zauważmy, że początkowa duża intensywność domykania wynika ze współuczestnictwa B i C w wielu – nawet ośmiu – niedomkniętych triadach typu 201. W miarę upływu czasu szansa na takie zjawisko maleje, bowiem użytkownicy zachęceni do nawiązania relacji poprzez wielu wspólnych znajomych na Instagramie, nawiązali je już wcześniej.

W uzupełnieniu powyższej oceny wizualnej wyników z Tab. 1, na Rys. 5 przedstawiono regresję liniową intensywności domykania i średniej liczmy wspólnych znajomych n w domykających się triadach. Zależność jest ewidentna, głównie w środku wykresu. Uzupełnia to naszą wiedzę o dynamice triad mało usieciowionych z rozdz. 3.3.

Tab. 1. Rozkład liczby wspólnych znajomych par węzłów domykających triady 201 w kolejnych migawkach

liczba n wspólnych znajomych	numer migawki											
	3	4	5	6	7	8	9	10	11	12		
11			11	11								
10												
9												
8	8											
7	4											
6	12	9	5					8				
5								4				
4		2		4	2							
3	16	11	2	9	9		5	18	3			
2	20	23	21	52	26	8	24	4				12
1	77	92	124	250	84	54	147	79	47	42		

Do wyznaczonej zależności należy podchodzić z rezerwą, gdyż zawiera ona wiele anomalii, które trudno uzasadnić – jak np. domknięcie się mocno usieciowionych triad dopiero w migawce 5 i 6 (niewidoczne po uśrednieniu przez regresję)., Zastanawia również niezmiennie wysoki udział we wszystkich migawkach konfiguracji, w których B i C mają 2 wspólnych znajomych: wydaje się, że taki fakt w ogóle się nie liczy. Natomiast, zanikające w czasie, uczestniczenie B i C w więcej niż 5 triadach jednocześnie może jedynie być objawem bardziej złożonych zjawisk sieciowych.



Rys. 5. Zależność intensywności domykania triad od aktualnej liczby wspólnych znajomych, tj. liczby triad 201 (por. grafika w lewym górnym rogu), w której uczestniczy domykające się łącze B-C.

3.4. Przypadki graniczne

Obserwacje poczynione w rozdz. 3.2 wykazały monotoniczny spadek intensywności słabo usieciowionych w funkcji czasu. Można postawić sobie dalsze pytanie, czy szybkość zaniku tej intensywności zależy od jakichkolwiek, obserwowalnych w posiadanych zbiorze danych, czynników? Wykonano zatem analizę rozkładów stopni wierzchołków A, B i C domykających się triad 201, ale tylko tych, w których B i C mają jeden wspólny wierzchołek A (por. ostatni wiersz w Tab. 1). Nie zaobserwowano jednak żadnej istotnej rozbieżności w kształtach tych rozkładów w początkowych i końcowych migawkach. Natomiast macierz współczynników korelacji różnych cech dodatkowych węzłów A, B i C wykazała niewielki dodatni wpływ stopnia wierzchołka B oraz niewielki ujemny wpływ stopnia wierzchołka C na intensywność domykania. Współczynnik korelacji intensywności i różnicy stopni B i C wyniósł 0,63. Taka zależność nie znajduje jednak logicznego wytłumaczenia w obrębie posiadanych i wziętych pod uwagę informacji o sieci.

Innym granicznym przypadkiem jest niedomknięcie się triad typu 201 w ogóle, przez cały czas trwania eksperymentu. Takich niedomykających się triad jest w sieci olbrzymia większość, wg [10] aż 99,6%. Porównano zatem rozkłady stopni wierzchołków A, B i C oraz ich sum (z ew. rozróżnieniem na stopnie wejściowe i wyjściowe) dla triad domykających się i nie. Również tutaj otrzymane wyniki są nierozstrzygujące.

4. PODSUMOWANIE

W pracy podjęto się wskazania czynników determinujących dynamikę najistotniejszego procesu w sieciach widzianych z perspektywy obecnych w nich triad: transformacji triad typu 201, z nieobecny połączeniem pomiędzy dwoma węzłami, w triady typu 300, z połączeniem nawiązanym. Z uwagi na zróżnicowaną i nieznaną historię triad zaobserwowanych podczas pierwszego pobierania danych z serwisu Instagram, analiza ma sens tylko w odniesieniu do triad nowych, pojawiających się w następnej migawce.

Na intensywność domykania ma wpływ usieciowienie triad 201 – rozumiane zarówno jako stopień poszczególnych węzłów triady, jak również jako liczba triad, do których należą węzły nawiązujące relację domykającą triadę. Większe usieciowienie skutkuje większym prawdopodobieństwem wczesnego domknięcia się triady, jednakże dla triad bardzo mocno usieciowionych na zależność tę zdają się wpływać inne czynniki, niewytłumaczalne posiadanymi lub wziętymi pod uwagę w analizie danymi. W przyszłości należy rozważyć bezwzględnie analizę szerszego otoczenia domykającej się triady, w celu znalezienia ewentualnych prawidłowości. Podejście takie zostało z powodzeniem zastosowane np. w wyjaśnianiu przyczyn rezygnacji abonentów z usług telekomunikacyjnych (*churn*) [8].

Triady mniej usieciowione domykają się wolniej, w naturalnym, trudnym do zaburzenia czynnikami zewnętrznymi, tempie. Wykonana analiza nie wykazała ewidentnego wpływu dodatkowych czynników na to tempo; faktem jest, że w rozważaniach ograniczono się wyłącznie do analizy stopni poszczególnych wierzchołków triady. Niewątpliwie, uzupełnienie posiadanych danych nowymi informacjami z serwisu Instagram (np. płcią i wiekiem użytkowników) mogłoby otworzyć nowe możliwości modelowania.

PODZIĘKOWANIA

Autor pragnie wyrazić gorące podziękowania panu Jakubowi Jarzyńskiemu, autorowi pracy [10] za współpracę i zaangażowanie w trakcie realizacji pracy magisterskiej, a następnie za udostępnienie pobranych danych z serwisu Instagram.

LITERATURA

- [1] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.
- [2] Bearman, P. S., & Moody, J. (2004). Suicide and friendships among American adolescents. *American journal of public health*, 94(1), 89-95.
- [3] Faust, K. (2006). Comparing social networks: size, density, and local structure. *Metodoloski zvezki*, 3(2), 185.
- [4] Faust, K. (2008). Triadic configurations in limited choice sociometric networks: Empirical and theoretical results. *Social Networks*, 30(4), 273-282.
- [5] Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395), 832-842.
- [6] GlobalWebIndex's quarterly report on the latest trends in social networking, http://www.globalwebindex.net/hubfs/Reports/GWI_Social_Report_-_Q3_2015_Summary.pdf (dostęp 2016.05.25)
- [7] Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360-1380.
- [8] W. Gruszczyski, P. Arabas. (2015) Wykorzystanie technik sieci społecznych w redukcji odejść klientów sieci telekomunikacyjnej, *Przegląd Telekomunikacyjny*, nr 8-9.
- [9] Holland, P. W., & Leinhardt, S. (1976). Local structure in social networks. *Sociological methodology*, 7(1), 1-45.
- [10] Jarzyński, Jakub (2016). Analiza triad w serwisach społecznościowych. Praca magisterska złożona na Wydziale Elektroniki i Technik Informacyjnych, Politechnika Warszawska.
- [11] Kolpakov, A. V., & McGuire, M. (2010). What Can Triadic Census Tell Us about the Development of Public Management Networks?, http://opensiuc.lib.siu.edu/pnconfs_2010/44/ (dostęp 2016.05.25).
- [12] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Sherrill, S., Ayzenshtat, I., ... & Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663), 1538-1542.
- [13] Socialpress, <http://socialpress.pl/2015/09/ponad-3-mln-polskich-uzytownikow-na-instagramie-wiekszosc-to-kobiety-i-nastolatki/> (dostęp 2016.05.25).