Załącznik 3 - Autoreferat

1. Wykształcenie

— **Stanford University**, 21.10.2013–20.12.2013, Stanford Center for Professional Development, Science Management and Commercialization. Staż badawczo-szkoleniowy w ramach programu MNiSW Top500 Innovators.
— **Politechnika Wrocławska**, 01.10.2008–30.10.2012, Wydział Informatyki i Zarządzania, studia doktoranckie w dziedzinie Informatyka. Tytuł rozprawy: *A Method for Group Extraction and Analysis in Multi-layered Social Networks*. **Praca obroniona z wyróżnieniem**.
— **The International Project Management Association (IPMA)**, certyfikacja z zarządzania projektem IMPA Level D (12.2012) certyfikat numer 635/2012, oraz certyfikat z zarządzania projektem badawczo - rozwojowym IMPA Level 4RD (03.2016) certyfikat numer 62/2016/RD. 11.2017 odnowienie obu certyfikatów na kolejne 5 lat.
— **Politechnika Wrocławska**, 01.02.2012–01.12.2012, Wydział Informatyki i Zarządzania, *Studia podyplomowe Zarządzanie projektem badawczym i komercjalizacja wyników badań.*
— **Blekinge Institute of Technology**, 01.09.2007–27.06.2012, degree of Master of Science, main field of study Computer Science. Praca magisterska *Key User Extraction Based on Telecommunication data.*
— **Politechnika Wrocławska**, 01.10 2003–07.07.2008, Wydział Informatyki i Zarządzania, Informatyka. Praca magisterka: *Key User Extraction Based on Telecommunication data.*
— **I Liceum Ogólnokształcące imienia Stefana Żeromskiego w Jeleniej Górze**, 01.09.1999–01.06.2003, profil matematyczno—fizyczno–informatyczny.

2. Informacje o dotychczasowym zatrudnieniu

— **University of Technology Sydney**, 26.09.2018–31.01.2019, Visiting Professor, Faculty of Engineering and Information Technology
— **Global Arena Research Institute**, Praga, 01.09.2018–obecnie, Stanowisko: Researcher
— **Zachodniopomorski Uniwersytet Technologiczny w Szczecinie**, 08.02.2017–obecnie, Stanowisko: wykonawca projektu
— **Unia Europejska/Komisja Europejska**, 22.01.2015–obecnie, Stanowisko: Ekspert/Recenzent
— **Narodowe Centrum Badań i Rozwoju**, 07.01.2014–obecnie, Stanowisko: Ekspert/Recenzent
— **Politechnika Wrocławska**, Wydział Informatyki i Zarządzania, Katedra Inteligencji Obliczeniowej, 15.11.2012–obecnie, Stanowisko: Adiunkt
— **Politechnika Wrocławska**, Wydział Informatyki i Zarządzania, Instytut Informatyki, 01.12.2010 – 15.11.2012, Stanowisko: Asystent
— **Research & Engineering Center**, 04.01.2010–31.12.2010, Stanowisko: Młodszy specjalista ds. oprogramowania, praca przy projekcie B+R finansowanym przez Narodowe Centrum Badań i Rozwoju

PB

— **Value Based Advisors Sp. z.o.o.**, 01.08.2008–30.11.2009, Stanowisko: Programista .NET

3. Główne osiągnięcie naukowe

Jako osiągnięcie naukowe pt. *Metody obliczeniowe i algorytmy do analizy sieci złożonych* przedkładam wymienione poniżej 11 prac. Informacje dodatkowe:

— Wszystkie przedstawione poniżej prace powstały w ramach realizacji projektów naukowych i/lub współpracy międzynarodowej, w związku z powyższym kolejność autorów nie zawsze odzwierciedla wkład w publikację. Przykładowo w publikacji (*"Predicting community evolution in social networks"* 2015) jestem dopiero trzecim autorem, jednakże mój wkład jest prawie dwa razy większy niż któregokolwiek z pozostałych autorów. W związku z powyższym dla każdej publikacji podałem listę autorów posortowaną po procentowym wkładzie w publikację.
— Deklaracje współautorów określające ich wkład w poszczególne publikacje stanowią Załącznik nr 5 do niniejszego wniosku o przeprowadzenie postępowania habilitacyjnego. Dodatkowo opis wkładu znajduje się w treści większości artykułów.
— Mój wkład jest scharakteryzowany krótko przy każdej pracy.
— Liczba cytowań przedstawia liczbę cytowań na dzień 16.04.2019 w serwisie Google Scholar, bez autocytowań to jest: wykluczono te prace cytujące, w których pojawia się chociaż jeden z autorów pracy cytowanej.
— Zarówno punkty MNiSW jak i Impact Factor wskazano na podstawie danych z serwisu Biblioteki Politechniki Wrocławskiej - DONA `https://dona.pwr.edu.pl/szukaj/`
— Pełen tekst wszystkich poniższych prac można pobrać pod adresem: `http://piotrbrodka.pl`
— Większość moich publikacji można pobrać ze strony ResearchGate: `https://www.researchgate.net/profile/Piotr_Brodka2`

3.1. Publikacje wchodzące w skład głównego osiągnięcia naukowego

1. **Piotr Bródka**, Anna Chmiel, Matteo Magnani, and Giancarlo Ragozini. "Quantifying layer similarity in multiplex networks: a systematic study". In: *Royal Society open science* 5.8 (2018), p. 171747
   — *Mój wkład:* [40%], stworzenie koncepcji badań, przygotowanie taksonomii, opracowanie miar, zaprojektowanie eksperymentów, zebranie i przygotowanie danych do badań, analiza wyników oraz przygotowanie manuskryptu.
   — *Wkład procentowy autorów:* **PB:40%**, ACh:35%, MM:15%, GR:10%
   — *Punkty IF:* 2,504,
   — *Punkty MNiSW:* 5,
   — *Cytowania*: 4.

2. Jarosław Jankowski, Boleslaw K Szymanski, Przemysław Kazienko, Radosław Michalski, and **Piotr Bródka**. "Probing Limits of Information Spread with Sequential Seeding". In: *Scientific reports* 8.1 (2018), p. 13996
   — *Mój wkład:* [20%], opracowanie metody i algorytmu, zaprojektowanie eksperymentu, opracowanie przykładowych symulacji i przygotowanie manuskryptu.
   — *Wkład procentowy autorów:* JJ:35%, RM:35%, **PB:20%**, BS:5%, PK:5%
   — *Punkty IF:* 4,122,
   — *Punkty MNiSW:* 40.

3. Belfin R.V., Grace Mary Kanaga E., and **Bródka Piotr**. "Overlapping community detection using superior seed set selection in social networks". In: *Computers & Electrical Engineering* 70 (2018),

pp. 1074–1083. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2018.03.012. URL: http://www.sciencedirect.com/science/article/pii/S0045790617318256
— *Mój wkład:* [25%], opracowanie algorytmów Superior Seed Set Selection oraz Superior Seed Set Expansion, zaprojektowanie eksperymentów, analiza wyników eksperymentu i opracowanie manuskryptu.
— *Wkład procentowy autorów:* GKE:40%, BRV:35%, **PB:25%**
— *Punkty IF:* 1,747,
— *Punkty MNiSW:* 20.

4. Jarosław Jankowski, **Piotr Bródka**, Przemysław Kazienko, Boleslaw K Szymanski, Radosław Michalski, and Tomasz Kajdanowicz. "Balancing speed and coverage by sequential seeding in complex networks". In: *Scientific reports* 7.1 (2017), p. 891
— *Mój wkład:* [30%], opracowanie koncepcji sekwencyjnego aktywowania wierzchołków początkowych, rozwinięcie koncepcji, zaprojektowanie eksperymentu, analiza wyników eksperymentu, przeprowadzenie dodatkowych eksperymentów, opracowanie manuskryptu;
— *Wkład procentowy autorów:* JJ:50%, **PB:30%**, PK:5%, RM:10%, BS:3%, TK:2%
— *Punkty IF:* 4,122,
— *Punkty MNiSW:* 40,
— *Cytowania*: 6.

5. Jarosław Jankowski, Radosław Michalski, and **Piotr Bródka**. "A multilayer network dataset of interaction and influence spreading in a virtual world". In: *Scientific data* 4 (2017), p. 170144
— *Mój wkład:* [33%], przygotowanie zbioru danych, analiza danych, tworzenie sieci, analiza sieci wielowarstwowej, analiza procesów rozprzestrzeniania i sieci, opracowanie manuskryptu.
— *Wkład procentowy autorów:* JJ:34%, **PB:33%**, RM:33%
— *Punkty IF:* 5,305,
— *Punkty MNiSW:* 5,
— *Cytowania*: 2.

6. Fredrik Erlandsson, **Piotr Bródka**, Martin Boldt, and Henric Johnson. "Do we really need to catch them all? A new User-guided Social Media Crawling method". In: *Entropy* 19.12 (2017), p. 686
— *Mój wkład:* [40%], opracowanie metody User-guided Social Media Crawling (USMC) oraz algorytmu Social Interaction Network Crawling Engine (SINCE), zaprojektowanie eksperymentów, analiza wyników oraz opracowanie manuskryptu.
— *Wkład procentowy autorów:* **PB:40%**, FE:40%, MB:15%, HJ:5%
— *Punkty IF:* 2,305,
— *Punkty MNiSW:* 30,
— *Cytowania*: 1.

7. Jarosław Jankowski, **Piotr Bródka**, and Juho Hamari. "A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world". In: *Behaviour & Information Technology* 35.11 (2016), pp. 926–945
— *Mój wkład:* [40%], badania nad wpływem warstw/typów interakcji na adopcję wirtualnego produktu i zaangażowanie użytkownika w dalsze jego rozprzestrzenianie, analiza pozostałych cech użytkowników oraz opracowanie manuskryptu.
— *Wkład procentowy autorów:* JJ:50%, **PB:40%**, JH:10%
— *Punkty IF:* 1,388,
— *Punkty MNiSW:* 25,
— *Cytowania*: 4.

PB

8. Fredrik Erlandsson, **Piotr Bródka**, Anton Borg, and Henric Johnson. "Finding influential users in social media using association rule learning". In: *Entropy* 18.5 (2016), p. 164
   — *Mój wkład:* [40%], opracowanie metody i algorytmu ARL, zaprojektowanie eksperymentów, analiza wyników oraz opracowanie manuskryptu.
   — *Wkład procentowy autorów:* **PB:40%**, FE:40%, AB:15%, HJ:5%
   — *Punkty IF:* 1,821,
   — *Punkty MNiSW:* 30,
   — *Cytowania*: 31.

9. Stanisław Saganowski, Bogdan Gliwa, **Piotr Bródka**, Anna Zygmunt, Przemysław Kazienko, and Jarosław Koźlak. "Predicting community evolution in social networks". In: *Entropy* 17.5 (2015), pp. 3053–3096
   — *Mój wkład:* [30%], autor metody, projektowanie eksperymentów, analiza wyników i opracowanie manuskryptu.
   — *Wkład procentowy autorów:* **PB:30%**, AZ:16,67%, BG:16,67%, JK:16,66, SS:15%, PK:5%
   — *Punkty IF:* 1,743,
   — *Punkty MNiSW:* 30,
   — *Cytowania*: 26.

10. Katarzyna Musial, **Piotr Bródka**, Przemysław Kazienko, and Jarosław Gaworecki. "Extraction of multilayered social networks from activity data". In: *The Scientific World Journal* 2014 (2014)
    — *Mój wkład:* [50%], opracowanie koncepcji i algorytmu do ekstrakcji sieci wielowarstwowych, zaprojektowanie eksperymentu, przeprowadzenie badań, analiza wyników badań oraz opracowanie manuskryptu.
    — *Wkład procentowy autorów:* **PB:50%**, KM:20%, PK:20%, JG:10%
    — *Punkty IF:* 1,219,
    — *Punkty MNiSW:* 30,
    — *Cytowania*: 2.

11. Radosław Michalski, Tomasz Kajdanowicz, **Piotr Bródka**, and Przemysław Kazienko. "Seed selection for spread of influence in social networks: Temporal vs. static approach". In: *New Generation Computing* 32.3-4 (2014), pp. 213–235
    — *Mój wkład:* [30%], opracowanie miar temporalnych, zaprojektowanie eksperymentu, analiza wyników badań oraz opracowanie manuskryptu.
    — *Wkład procentowy autorów:* TK:40%, **PB:30%**, RM:15%, PK:15%
    — *Punkty IF:* 0,821,
    — *Punkty MNiSW:* 25,
    — *Cytowania*: 18.

Łączną liczba punktów MNiSW, punktów Impact Factor oraz inne podstawowe informacje na temat głównego osiągnięcia naukowego zebrano w Tabeli 1.

3.2. Omówienie celu naukowego ww. prac i osiągniętych wyników.

Dwadzieścia lat temu wraz z pojawieniem się zbiorów danych umożliwiających odtworzenie struktury sieci oraz szybkim wzrostem mocy obliczeniowych sprzętu komputerowego narodziła się nowa gałąź nauki zwana nauką o sieciach (Network Science[1]). Obszar, który do tej pory był domeną fizyków,

[1] https://en.wikipedia.org/wiki/Network_science

PB

Tabela 1. Skrótowe podsumowane głównego osiągnięcia naukowego

Liczba publikacji	11
Liczba publikacji z największym wkładem	5
Liczba publikacji z wiodącym wkładem	10
Liczba publikacji ze znaczącym wkładem	11
Suma współczynnika Impact Factor	27,097
Suma punktów MNiSW	280
Liczba cytowań według Google Scholar bez autocytowań* (stan na 16.04.2019)	94

* – Wykluczono te prace cytujące, w których pojawia się chociaż jeden z autorów pracy cytowanej

matematyków i socjologów, stał się także domeną informatyków. Co więcej informatycy zaczeli odgrywać kluczową rolę w tym obszarze prowadząc do powstania takich dziedzin jak Obliczeniowa nauka o sieciach (Computational netowrk science[2],[3]) czy Obliczeniowe nauki społeczne (Computational social science[4]). Cel naukowy mojej pracy, opisany zarówno w tym rozdziale jak i w kolejnym, to pogłębianie światowej wiedzy na temat fenomenu jakim są sieci złożone, a w szczególności wprowadzenie istniejących bądź nowych, metod, algorytmów i narzędzi informatyki w celu umożliwienia lub ułatwienia analizy sieci złożonych. Droga, którą przebyłem od czasu otrzymania stopnia doktora uzmysłowiła mi, że jest do zrobienia więcej niż przypuszczałem, jednakże wierzę, że mój wkład w ten obszar jest znaczący (co po części jest odzwierciedlone w cytowaniach moich prac). Wkład ten można podzielić na pięć uzupełniających się i wzajemnie przenikających obszarów (Rysunek 1):

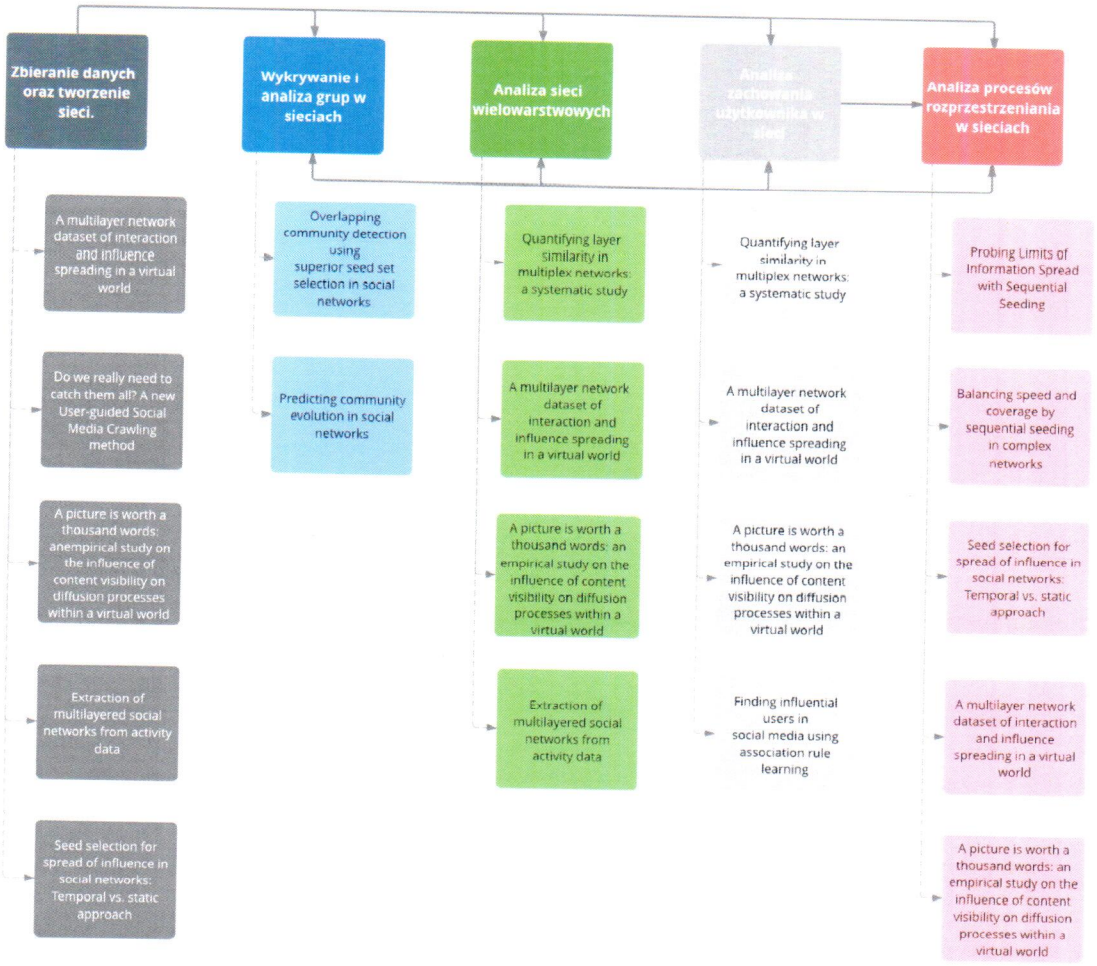
1. Metody automatycznej akwizycji rzeczywistych danych sieciowych oraz metody automatycznej ekstrakcji sieci różnego rodzaju
2. Metody i algorytmy do wykrywania i analizy grup w sieciach złożonych
3. Analiza sieci wielowarstwowych
4. Analiza zachowania i pozycji węzła (użytkownika) w procesach zachodzących w sieci
5. Analiza procesów rozprzestrzeniania w sieciach

Mój wkład w każdy z tych obszarów omówię w poniższych rozdziałach. W trakcie opisu, by ułatwić rozróżnienie pomiędzy pracami należącymi do cyklu a nienależącymi do cyklu, będę używał dwóch stylów odnoszenia się do moich prac. Styl pierwszy będzie się odnosił do prac z głównego osiągnięcia naukowego, przykład: (*"Quantifying layer similarity in multiplex networks: a systematic study"* 2018). Natomiast drugi, standardowy, styl będzie się odnosił do pozostałych moich publikacji nie wchodzących w skład głównego osiągnięcia naukowego, przykład: [3].

3.2.1. Metody automatycznej akwizycji rzeczywistych danych sieciowych oraz metody automatycznej ekstrakcji sieci różnego rodzaju

Nauka o sieciach (Network Science), a w szczególności Obliczeniowa nauka o sieciach (Computational Network Science), zawdzięczają, w dużej mierze, swoje powstanie i dynamiczny rozwój zwiększającej się dostępności danych na podstawie, których można badać interakcje pomiędzy wierzchołkami w sieci. Co prawda w dalszym ciągu większość badań prowadzi się na sieciach sztucznych, wygenerowanych według wybranych modeli (np. Erdos-Renyi, Watts-Strogatz, Barabási–Albert, itd.), ale odsetek prac wykorzystujących dane rzeczywiste nieustannie wzrasta. Osoba, chcąca oprzeć swoje badania na danych rzeczywistych zazwyczaj ma do wyboru:

[2] https://www.sciencedirect.com/book/9780128008911/computational-network-science
[3] https://www.computationalnetworkscience.org/
[4] https://en.wikipedia.org/wiki/Computational_social_science



Rysunek 1. Poglądowy schemat przedstawiający publikacje w ramach głównego osiągnięcia naukowego wraz z przypisaniem do poszczególnych obszarów

1. pobrać dane przygotowane przez inne osoby i udostępnione w repozytorium, np. Network Repository[5], niestety wtedy nie zawsze wiemy jak dane zostały zebrane, czy są kompletne, co zostało z nich usunięte itd.;
2. kupić dane (np. Twitter's enterprise API) co wiąże się zazwyczaj z dużymi kosztami, dodatkowo wiele firm (np. Facebook) nie sprzedaje bezpośrednio swoich danych;
3. wynegocjować darmowy dostęp do danych do celów naukowych;
4. samodzielnie zebrać dane poprzez tradycyjne podejścia znane z socjologi jak ankiety czy obserwacje (problem z kompletnością danych, reprezentatywnością danych itd.), lub użyć metod do automatycznego zbierania treści np. przy pomocy robotów internetowych (problemy: zmienne API, kwestie prawne, szybszy przyrost danych niż jesteśmy w stanie je zbierać).

To ostatnie podejście czyli automatyczne zbieranie danych przy pomocy robotów internetowych oraz problem szybkiego przyrostu danych został zaadresowany w pracy (*"Do we really need to catch them all? A new User-guided Social Media Crawling method"* 2017). Zespół z Blekinge Institute of Technology zajmował się zbieraniem danych z publicznych stron portalu Facebook (czyli stron, które

[5] `http://networkrepository.com`

są publicznie dostępne i nie wymagają logowania się do serwisu Facebook np. `https://www.facebook.com/cnninternational/`) od 2012 roku. Ponieważ strony publiczne stanowią niewielki odsetek wszystkich stron na portalu Facbook, problem szybszego przyrostu treści w stosunku do możliwości robota napotkali krótko przed moim dołączeniem do zespołu (więcej na ten temat w sekcji 3.2.4), jednak dalej byli w stanie zbierać dane dla kilkuset wybranych stron. By zaadresować ten problem wspólnie z doktorantem Fredrikiem Erlandssonem opracowaliśmy metodę i algorytm User-guided Social Media Crawling (USMC). Działanie metody polega na istnieniu dwóch robotów internetowych. Pierwszy "szybki" robot zbiera tylko ogólne dane na temat tego jakie posty znajdują się na danej stronie, kiedy każdy z nich powstał, ile ma polubień (likes) oraz ile komentarzy. Te metadane pozwalają nam określić, które posty są interesujące dla użytkowników, które posty użytkownicy lubią i w dyskusji pod którymi postami użytkownicy biorą udział. Dzięki temu tak naprawdę pozwalamy użytkownikom serwisu wskazać te posty, które są wartościowe, i które, przy ograniczonych zasobach jakie mamy, powinniśmy zebrać w pierwszej kolejności. Na podstawie złożenia tych metadanych drugi "dokładny" robot pobiera posty na stronie, zaczynając od najważniejszego w rankingu, aż do osiągnięcia pożądanej wielkości próbki lub do wyczerpania się z góry zadanego czasu na daną stronę. Przeprowadzone badania, przy wykorzystaniu danych zebranych ze 160 publicznych stron portalu Facebook (jeden z największych istniejących poza firmą Facebook zbiorów danych), pokazały, że średnio możemy zgromadzić 75% wszystkich interakcji pomiędzy użytkownikami poprzez zebranie zaledwie 20% postów, natomiast czas zebrania tych 75% interakcji jest o 53% krótszy niż w przypadku użycia dotychczasowego podejścia. W kolejnym kroku dla każdej strony zbudowaliśmy sieci przy użyciu wszystkich danych oraz częściowych zbiorów zawierających X% najważniejszych postów i porównaliśmy te sieci. Okazało się, że dzięki zebraniu zaledwie 20% najważniejszych postów możemy stworzyć sieć, która zawiera 75% wszystkich wierzchołków i 85% wszystkich krawędzi a co ważniejsze zachowany zostaje rozkład stopni węzłów w sieci. Szczegółowe wyniki dla każdej ze 160 stron można znaleźć w suplemencie artykułu (*"Do we really need to catch them all? A new User-guided Social Media Crawling method"* 2017).

Badania nad procesami rozprzestrzeniania w sieciach wielowarstwowych prowadzonych w ramach publikacji (*"A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world"* 2016) (szerszy opis prac w ramach tej publikacji jest w sekcji 3.2.4), pokazały, że nie istnieje na świecie zbiór danych, który zawierałby rzeczywiste procesy rozprzestrzeniania na rzeczywistej sieci wielowarstwowej. W związku z tym, część danych użytych w ramach publikacji (*"A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world"* 2016) pozyskana z polskiej gry Timik.pl została po dodatkowym przetworzeniu i anonimizacji udostępniona w repozytorium Harvard Dataverse[6], natomiast pełna analiza tego zbioru danych, sieci wielowarstwowej utworzonej na podstawie tych danych oraz procesów rozprzestrzeniania została umieszczona w publikacji (*"A multilayer network dataset of interaction and influence spreading in a virtual world"* 2017).

W powyższych opisach nie poświęcam za wiele uwagi etapowi ekstrakcji sieci społecznej z posiadanych danych co może tworzyć wrażenie że jest to proces szybki i prosty. Jednakże zazwyczaj tak nie jest i decyzja o tym co ma być wierzchołkiem a co krawędzią oraz sam proces tworzenia sieci jest nietrywialny, w szczególności w przypadku sieci wielowarstwowych (*"Extraction of multilayered social networks from activity data"* 2014) i dynamicznych (*"Seed selection for spread of influence in social networks: Temporal vs. static approach"* 2014). W pracy (*"Extraction of multilayered social networks from activity data"* 2014), na przykładzie stosunkowo prostych danych pochodzących z forum internetowego opracowaliśmy algorytmy pozwalające tworzyć różne sieci wielowarstwowe w zależności od tego jaki sposób modelowania relacji pomiędzy poszczególnymi obiektami zastosujemy. Co więcej, te różne sposoby sposoby modelowania pozwalają na odkrycie nowych relacji pomiędzy obiektami w sieci,

[6] Jankowski, Jarosław; Radosław Michalski; Piotr Bródka, 2017, "Spreading processes in multilayer complex network within virtual world", https://doi.org/10.7910/DVN/V6AJRV, Harvard Dataverse, V1

które nie są oczywiste i mogłyby być pominięte w przypadku zastosowania standardowych podejść do ekstrakcji sieci.

Oczywiście we wszystkich pozostałych pracach także odbywał się etap akwizycji danych i ekstrakcji sieci złożonej, jednakże, nie był to najważniejszy ani nawet jeden z głównych elementów dlatego też te prace opisano w kolejnych sekcjach.

3.2.2. Metody i algorytmy do wykrywania i analizy grup w sieciach złożonych

W ramach doktoratu opracowałem metodę do ekstrakcji historii grupy (GED - Group Evolution Discovery) [5], a jako przykład jej wykorzystania zaproponowałem próbę predykcji tego co się stanie z grupą/społecznością w przyszłości. Następnie wspólnie z doktorantem (obecnie dr inż. Stanisław Saganowski), którego byłem promotorem pomocniczym podjęliśmy się sprawdzenia tej teorii w ramach jego doktoratu. Dla każdego z trzech różnych zbiorów danych (dane z serwisu DBLP, serwisu Facebook i serwisu Salon24), dokonaliśmy podziału na okresy. Następnie, dla każdego z nich stworzyliśmy sieć złożoną przestawiającą interakcje pomiędzy użytkownikami w tym okresie. Kolejnym krokiem była ekstrakcja grup w każdej sieci przy użyciu metody CPM[7], oraz ekstrakcja historii zmian (ewolucji) grupy przy użyciu metody GED [5] oraz SGCI [13] (opracowanej po doktoracie wspólnie z naukowcami z AGH). Każdą z grup w każdym momencie jej ewolucji opisywaliśmy cechami, 29 (dla metody SGCI) lub 31 (dla medoty GEP), które charakteryzowały jej stan w danym okresie. Na tak przygotowanych danych przeprowadzaliśmy klasyfikację tego co się stanie z grupą w przyszłości, przy użyciu czterech różnych podejść (C4.5 decision tree, Random forest, Adaptive Boosting, Bootstrap aggregating), sprawdzaliśmy jakie cechy mają największy wpływ na jakość klasyfikacji, oraz jak długą historię grupy (w postaci liczy poprzednich stanów) musimy znać by mieć jak najlepsze wyniki klasyfikacji. Wyniki pokazały, że najlepsze rezultaty uzyskujemy dla historii o długości siedem (siedem poprzednich stanów), chociaż klasyfikatory używają głównie cech z trzech ostatnich stanów grupy. Najlepszymi metodami okazały się RandomForest oraz AdaBoost. Szczegółowe wyniki zostały opisane w publikacji (*"Predicting community evolution in social networks"* 2015).

W kolejnych latach badania te były kontynuowane w celu przebadania większej liczby klasyfikatorów (11), metod ekstrakcji grup, liczby cech (ponad 80), większej liczby zbiorów danych itd. Wszystkie te prace doprowadziły do powstania metody GEP (Group Evolution Prediction) oraz zostały zebrane w pracy doktorskiej dr inż. Stanisława Saganowskiego, która otrzymała wyróżnienie, oraz publikacji [33], która jest aktualnie recenzowana.

Drugi kierunek badań w obszarze grup w sieciach złożonych wywodzi się z moich prac nad procesami rozprzestrzeniania w sieciach złożonych i analizą istotności węzła w sieci opisanych szerzej w sekcji 3.2.5 oraz 3.2.4. W ramach tych prac poswatał pomysł by sprawdzić czy możliwe jest wykorzystanie metod do wykrywania węzłów początkowych (seeds), które inicjują proces rozprzestrzeniania w sieci, jako węzłów/liderów wokół których będzie można zbudować grupy w sieci. Badania rozpocząłem razem z doktorantem z Indii, który odbył u mnie dwumiesięczny staż naukowy, oraz jego promotorem. Opracowaliśmy algorytm o nazwie Superior Seed Set Selection (4-S), która używa złączenia czterech miar pozycji węzła w sieci złożonej (Degree centrality, PageRank centrality, Local Clustering Coefficient oraz Eigenvector centrality) do wyznaczenia węzłów wokół, których będą budowane grupy. Następnie opracowaliśmy metodę i algorytm o nazwie Superior Seed Set Expansion, który na podstawie wcześniej wyznaczonego przez metodę 4-S zbioru buduje grupy. Całe rozwiązanie zostało przebadane na trzech rzeczywistych, benchmarkingowych zbiorach danych oraz porównane do pięciu innych algorytmów do ekstrakcji grup. Otrzymane wyniki potwierdziły, że algorytm poprawnie identyfikuje grupy i jest od 5 do 10 procent szybszy od najszybszego algorytmu z grupy algorytmów do których go porównywaliśmy. Szczegółowy opis prac oraz otrzymanych wyników znajduje się w publikacji (*"Overlapping community detection using superior seed set selection in social networks"* 2018). Aktualnie trwają prace nad mo-

[7] http://www.cfinder.org/

PB

dyfikacją metody 4-S tak by przebadać inne kombinacje miar określających pozycję węzła w sieci, oraz nad rozszerzeniem obu algorytmów tak by działały także na sieciach wielowarstwowych.

3.2.3. Analiza sieci wielowarstwowych

Pierwszym etapem analizy sieci wielowarstwowych jest pozyskanie danych i tworzenie na ich podstawie sieci wielowarstwowej co zostało opisane w pracach (*"Extraction of multilayered social networks from activity data"* 2014) oraz (*"A multilayer network dataset of interaction and influence spreading in a virtual world"* 2017) (Sekcja 3.2.1). Duża część moich publikacji w taki czy inny sposób zawiera analizę sieci wielowarstwowej. Jednakże zazwyczaj nie jest to główny przedmiot badań. Podobnie też miało być gdy rozpoczynaliśmy prace badawcze z dr Anną Chmiel i prof Matteo Magnaim. Początkowo miały być to badania odpowiadające na pytanie czy na podobnych warstwach proces rozprzestrzeniania wygląda tak samo. Jednak gdy zaczęliśmy zgłębiać problem określania podobieństwa pomiędzy warstwami, okazało się że pomimo wielu prac na ten temat nie ma ujednoliconej nomenklatury w tym obszarze. Różni autorzy nazywają te same miary w różny sposób i badają kilka razy tą samą miarę co powoduje, że tak naprawdę niewiele z nich zostało sprawdzonych. W związku z powyższym zgodnie z metodyką, Design Thinking, przedefiniowaliśmy problem i zajęliśmy się problemem określania podobieństwa pomiędzy warstwami. Pierwszym elementem było stworzenie jednolitej taksonomii w tym obszarze. W ramach niej zaproponowaliśmy macierz cech (property matrix) na podstawie, której możemy policzyć dowolne podobieństwa cech dwóch warstw, niezależnie czy są to cechy wierzchołka, krawędzi, motywy, grupy czy dowolnej innej struktury w sieci. Dodatkowo dokonaliśmy przeglądu istniejących miar, ujednoliciliśmy nazewnictwo oraz zaproponowaliśmy szereg nowych miar. Następnie, przy użyciu 23 rzeczywistych sieci wielowarstwowych dokonaliśmy analizy istniejących i nowych miar (łącznie 50). Na podstawie uzyskanych wyników opracowaliśmy zestaw wytycznych w jakich warunkach i dla sieci o jakiej charakterystyce możemy używać danej miary a kiedy nie, przykładowo miary Hammana i Simple Matching Coefficient nie mogą być stosowane dla sieci rzadkich czyli de facto prawie wszystkich sieci rzeczywistych. Dodatkowo przeprowadziliśmy analizę korelacji pomiędzy poszczególnymi miarami i wskazaliśmy te które można stosować zamiennie dla określonych typów sieci. Szczegółowe wyniki oraz wszystkie wytyczne są opisane w pracy (*"Quantifying layer similarity in multiplex networks: a systematic study"* 2018).

Sieć wielowarstwowa oraz wpływ warstw na proces rozprzestrzeniania treści był analizowany także w publikacji (*"A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world"* 2016), jednakże głównym przedmiotem badań były cechy i zachowania użytkownika oraz ich wpływ na proces rozprzestrzeniania dlatego też szerszy opis tych badań umieszczono w sekcji 3.2.4.

Chciałbym też, dodać, że sieci temporalne, które były przedmiotem badań w pracach (*"Seed selection for spread of influence in social networks: Temporal vs. static approach"* 2014) oraz (*"Predicting community evolution in social networks"* 2015) to, de facto, także sieci wielowarstwowe gdzie kolejne warstwy zawierają interakcje, które odbyły się w danym okresie. Trzeba jednak pamiętać, że w tym przypadku ważna jest kolejność warstw gdyż reprezentuje ona czas dlatego zazwyczaj analizuje się ten obszar osobno.

3.2.4. Analiza zachowania i pozycji węzła (użytkownika) w procesach zachodzących w sieci

Od analizy zachowania i pozycji węzła w sieci w ramach pracy magisterskiej[8] dla British Telecom zaczęła się moja przygoda ze światem sieci złożonych. Po pracy magisterskiej w czasie pierwszego roku doktoratu jeszcze kontynuowałem ten obszar ale dość szybko go zarzuciłem by skupić się na doktoracie. We wrześniu 2015 roku w czasie organizacji konferencji ENIC 2015[9] na Blekinge Institute of Technology spotkałem profesora Henrica Johnsona, którego znałem już wcześnie z racji moich studiów na tej

[8] https://arxiv.org/abs/1302.1369
[9] https://enic.cse.bth.se/

PB

uczelni oraz wspólnej publikacji w temacie bezpieczeństwa w sieciach [6]. W czasie rozmowy okazało się, że jego doktorant pomimo posiadania olbrzymiego zbioru danych zebranego z portalu Facebook nie potrafi przeanalizować zachowań użytkowników i określić ich pozycji w sieci (głownie z powodu wielkości zbioru). Dzieki tej dyskusji powstał pomysł na nowy sposób określania pozycji, który został opublikowany w pracy [7], zostałem zaproszony do bycia co-promotorem doktoranta oraz rozpoczęliśmy długotrwałą i owocną współpracę naukową. Pierwszym elementem tej współpracy była właśnie wcześniej wspomniana nowa metoda identyfikowania kluczowych użytkowników w sieci przy użyciu reguł asocjacyjnych. Dzięki wykorzystaniu reguł asocjacyjnych i prostego założenia, że osoby, które często zaczynają dyskusję, w której następnie bierze udział wiele innych osób, są osobami ważnymi dla danej społeczności, opracowaliśmy nową metodę i algorytm ARL. Potrafi on znaleźć kluczowe osoby na "czystych" danych bez potrzeby projekcji relacji użytkowników w stosunku do obiektów (posty i komentarze) do sieci społecznej, której to sieci potrzebujemy by użyć "tradycyjnych" metod wykrywania użytkowników kluczowych takich jak stopień węzła czy PageRank. Przeprowadzone badania pokazały, że nie ma statystycznie istotnej różnicy pomiędzy wynikami osiąganymi przez ARL a PageRank, a dzięki pominięciu kosztownego procesu projekcji sieci, ARL jest średnio 36 razy szybszy niż stopień węzła i 70 razy szybszy niż PageRank (badania przeprowadzono na 108 różnych zbiorach danych pochodzących z publicznych stron portalu Facebook). W ramach badań dokonaliśmy jeszcze jednego interesującego odkrycia. Okazało się, że 10% i 20% najważniejszych użytkowników w sieci wytwarza odpowiednio 82% i 96% treści na stroni. Szczegółowy opis i wyniki badań znajdują się w pracy (*"Finding influential users in social media using association rule learning"* 2016). Algorytm ARL wykorzystano także w późniejszych badaniach nad detekcją węzłów początkowych dla procesu rozprzestrzeniania w sieciach wielowarstwowych [9].

Pracą, łączącą prawie wszystkie obszary, którymi zajmowałem się po doktoracie w ramach głównego osiągnięcia naukowego jest (*"A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world"* 2016). W ramach tej pracy sprawdaliśmy, jakie cechy wpływały na adopcję nowych przedmiotów i udział graczy w ich dalszym rozprzestrzenianiu. W czasie istnienia gry i portalu Timik.pl, jej właściciele w celu zwiększenia zainteresowania serwisem i zwiększenia sprzedaży kont premium rozpoczęli pięć różnych kampanii rozprzestrzeniania darmowych przedmiotów wśród graczy. Po zakończeniu kampanii chcieli się dowiedzieć dlaczego jedne kampanie miały większy zasięg i zaangażowały większą liczbę użytkowników a inne mniejszą. Utworzyliśmy i przeanalizowaliśmy 15 różnych cech, na które składały się cechy użytkownika, cechy poszczególnych warstw sieci wielowarstwowej oraz cechy samych kampanii. Z pośród nich tylko jedna cecha miała zawsze pozytywny wpływ na adopcję przedmiotu i aktywny udział w jego dalszym rozprzestrzenianiu a była to możliwość wcześniejszego zobaczenia produktu gdy używa go inny użytkownik. Pozostałe cechy miały wpływ na kampanię ale w mniejszym zakresie lub w określonych warunkach. Przykładowo gdy mechanika procesu rozprzestrzeniania była "łatwa" (by przekazać przedmiot wystarczyło by użytkownik kliknął innego użytkownika na ekranie) to kampanie miały większy zasięg (przedmiot docierał do większej liczby użytkowników) ale adopcja (rozpoczęcie używania przedmiotu, które często prowadzi do wykupienia konta premium) była niska. Natomiast gdy mechanika procesu rozprzestrzeniania była "trudna" (by przekazać przedmiot trzeba było być przyjacielem drugiego użytkownika oraz wysłać mu specjalną wiadomość z przedmiotem) to zasięg kampanii był nawet pięć razy mniejszy niż w przypadku "łatwej kampanii". Natomiast odsetek osób, które zaadoptowały nowy produkt była nawet cztery razy wyższa niż w przypadku "łatwej kampanii". To może oznaczać, że gdy mamy trudniejszy mechanizm to wysyłający mniej się angażują, ale jak już ktoś otrzyma nowy przedmiot od przyjaciela to z większym prawdopodobieństwem go użyje. Szczegółowa analiza wszystkich charakterystyk jest przedstawiona w pracy (*"A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world"* 2016). Wyniki tych prac mogą być w łatwy sposób wykorzystane do projektowania różnych kampanii w innych grach typu "wirtualny świat". Co więcej, wyniki naszych prac zainteresowały prof. Cuihua Shen z University of California

PB

Davis, która zajmuje się badaniem mechanizmów społecznych zachodzących w grach komputerowych i aktualnie prowadzimy wspólne badania na identyfikacją cech, które mają wpływ na zakup produktów premium w grach typu "wirtualny świat". Pierwsze wyniki prac były prezentowane na konferencjach SUNBELT 2018 (Amsteradam, Holandia) oraz IC2S2 2018 (Evanston, USA).

3.2.5. Analiza procesów rozprzestrzeniania w sieciach

W niedługim czasie po uzyskaniu stopnia doktora, rozpocząłem badania nad procesami rozprzestrzeniania w sieciach. Skupiałem się głównie na problemach związanych ze zbiorem wierzchołków początkowych służących do inicjowania procesu rozprzestrzeniania co jest w pewnym sensie naturalną konsekwencja moich prac opisanych w sekcji 3.2.4. Te problemy to: wpływ sposobu tworzenia sieci na ten zbiór (*"Seed selection for spread of influence in social networks: Temporal vs. static approach"* 2014), sposób aktywacji wierzchołków początkowych (*"Balancing speed and coverage by sequential seeding in complex networks"* 2017), (*"Probing Limits of Information Spread with Sequential Seeding"* 2018) czy algorytmy do wyboru zbioru w sieciach wielowarstwowych [9]. Dodatkowo zagadnienie rozprzestrzeniania wpływu w sieciach wielowarstwowych jest tematem grantu finansowanego przez NCN (SONATA), którego jestem kierownikiem `http://www.multispread.pwr.edu.pl/`.

W pierwszych badaniach skupiłem się na sprawdzeniu jak sposób budowy sieci wpływa na zbiór wierzchołków początkowych a w konsekwencji na cały proces rozprzestrzeniania. Analizowaliśmy dwa scenariusze "tradycyjna" sieć statyczna zawierająca wszystkie interakcje oraz sieć dynamiczna podzielona na okna czasowe o różnej granularności zawierająca w danym oknie tylko interakcje, które miały miejsce w zadanym okresie. Opracowaliśmy metody wyboru zbioru wierzchołków początkowych dla sieci temporalnych, a następnie przeprowadziliśmy eksperymenty na pięciu rzeczywistych zbiorach danych oraz pięciu różnych metodach wyboru zbioru wierzchołków początkowych. Otrzymane rezultaty jasno pokazały, że używając sieci dynamicznej i uwzględniając w wyborze zbioru zmiany w sieci, jesteśmy w stanie wyznaczyć lepszy zbiór, co z kolei powoduje, że końcowa liczba aktywowanych węzłów jest dwa razy większa niż dla sieci statycznej. Szczegółowy opis przeprowadzonych badań i uzyskanych wyników znajduje się w pracy (*"Seed selection for spread of influence in social networks: Temporal vs. static approach"* 2014).

Następnie wspólnie z Prof. Jarosławem Jankowskim opracowaliśmy nowatorską koncepcję sekwencyjnego aktywowania wierzchołków początkowych. W uproszczeniu polega ona na tym by zamiast aktywować wszystkie wierzchołki ze zbioru początkowego na samym początku, jak to było robione do tej pory, aktywujemy tylko część z nich i obserwujemy jak przebiega proces rozprzestrzeniania. Natomiast, "zaoszczędzone wierzchołki" wykorzystujemy dopiero gdy proces spowolni lub się zatrzyma. Pierwsza publikacja w tym obszarze (*"Balancing speed and coverage by sequential seeding in complex networks"* 2017) skupiała się na przebadaniu tej koncepcji. Sprawdziliśmy pięć różnych strategii sekwencyjnego aktywowania wierzchołków początkowych, na 15 sieciach oraz przy wykorzystaniu 5 strategi wyboru zbioru wierzchołków początkowych oraz różnych parametrów samego procesu rozprzestrzeniania (łącznie 1875 różnych przypadków). Symulacje potwierdziły nasze przypuszczenia, w 85% przypadków sekwencyjne aktywowanie wierzchołków początkowych dawało lepsze rezultaty niż jednorazowa aktywacja. W kolejnej publikacji (*"Balancing speed and coverage by sequential seeding in complex networks"* 2017), skupiliśmy się na: (i) analitycznym udowodnieniu, że w tych samych warunkach sekwencyjna aktywacja wierzchołków, w najgorszym przypadku, da te same rezultaty co jednorazowa aktywacja; (ii) przeprowadzeniem badań także dla sieci skierowanych; (iii) pokazaniem, że sekwencyjna aktywacja wierzchołków nawet z prostą metodą wyboru zbioru początkowego wierzchołków (stopień węzła) jest lepsza od algorytmu zachłannego (greedy) z jednorazową aktywacją; oraz (iv) pokazaniem, jak blisko sekwencyjna aktywacja wierzchołków pozwala nam się zbliżyć do maksymalnego możliwego zasięgu procesu rozprzestrzeniania dla danej sieci. Udało nam się zrealizować wszystkie powyższe cele, a przeprowadzone badania na, siedmiu rzeczywistych sieciach i różnych parametrach procesów rozprzestrzeniania (łącznie 16,2 miliona symulacji), pokazały, że w 92% przypadków sekwencyjna aktywacja

węzłów ze stopniem węzła jest lepsza niż algorytm zachłanny z jednorazową aktywacją. Co więcej, sekwencyjna aktywacja pozwala na skrócenie dystansu pomiędzy algorytmem zachłannym (uznawanym do tej pory za najlepszy) a maksymalnym możliwym zasięgiem nawet o 83%. Prace nad sekwencyjnym aktywowaniem wierzchołków początkowych są dalej prowadzone a ich wyniki prezentowane są w innych publikacjach, które ze względu mniej znaczące odkrycia [16], [18] lub mniejszy mój wkład [23], nie zostały uwzględnione w głównym osiągnięciu naukowym.

Zganienie procesów rozprzestrzeniania poruszają też publikacje omówione wcześniej jak (*"A multilayer network dataset of interaction and influence spreading in a virtual world"* 2017) czy (*"A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world"* 2016).

Natomiast ja w ramach tego obszaru aktualnie skupiam się na badaniach nad sekwencyjnym aktywowaniem wierzchołków początkowych w sieciach wielowarstwowych, oraz nad wieloma współwystępującymi procesami rozprzestrzeniania w sieciach wielowarstwowych [3].

3.3. Aktywności wspierające główne osiągnięcie naukowe

Warto podkreślić, że osiągniecie naukowe zostało wypracowane w ramach długotrwałej strategii, której realizację rozpocząłem zaraz po doktoracie, a która miała na celu rozwój nauki o sieciach złożonych w Polsce i na świecicie. W ramach tej strategii zrealizowano:

- liczne publikacje, które zostały omówione w ramach głównego osiągnięcia naukowego ale także inne publikacje z tego obszaru, które ze względu na mniejszy wkład w ich powstanie, brak współczynnika Impact Factor, lub bycie jeszcze w recenzji, zostały ujęte w pozostałych publikacjach (sekcja 4.1.1 oraz 4.1.2)
- Dwanaście projektów finansowanych przez różne agencje z obszaru analizy sieci złożonych (sekcja 4.2)
- Organizację konferencji, warsztatów (sekcja 4.6) oraz edytorstwo książek i numerów specjalnych (sekcja 4.3)
- współpracę międzynarodową (sekcja 4.7), która jest odzwierciedlona zarówno w publikacjach, projektach jak i w stażach zagranicznych (sekcja 4.5)