Załącznik nr 3

Autoreferat

1 Imię i nazwisko

Mariusz Kamola

2 Uzyskane dyplomy i stopnie naukowe

2004 r. — stopień doktora nauk technicznych w specjalności automatyka uzyskany na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej za rozprawę pt. *Algorithms for Optimisation Problems with Implicit and Feasibility Constraints*.

1997 r. — tytuł magistra inżyniera w specjalności automatyka uzyskany w Instytucie Automatyki i Informatyki Stosowanej Politechniki Warszawskiej za pracę pt. *Sterowanie z powtarzaną optymalizacją. Optymalizacja z nierównomierną dyskretyzacją sterowania*.

3 Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych

2002–2004 — asystent w Instytucie Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska (WEiTI PW)

2002–2004 — asystent w Pionie Naukowym, Naukowa i Akademicka Sieć Komputerowa (NASK jbr)

2004–obecnie — adiunkt w Instytucie Automatyki i Informatyki Stosowanej, WEiTI PW

2004–obecnie — adiunkt w Centrum Badań i Transferu Technologii, NASK Państwowy Instytut Badawczy

4 Omówienie osiągnięcia naukowego

Jako osiągnięcie naukowe wskazuję cykl powiązanych tematycznie artykułów naukowych, zgodnie z art. 219 pkt. 1 ustawy z dn. 3 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce.

4.1 Tytuł osiągnięcia naukowego

Metody analizy i odkrywania struktury sieci społecznych i technologicznych

4.2 Lista publikacji składających się na osiągnięcie naukowe

Poniżej podaję listę publikacji składających się na osiągnięcie naukowe w kolejności ich występowania w opisie osiągnięcia. Każdej pozycji towarzyszy krótki opis dzieła, informacja o moim udziale oraz wskaźniki bibliometryczne: liczba punktów MNiSW w roku wydania pracy, *Impact Factor*, liczba cytowań w Web of Science (WoS), Scopus i Google Scholar (GS).

[H1] M. Kamola (70%), P. Arabas, *Sieci społeczne i technologiczne. Jak zrozumieć, jak wykorzystać,* Wydawnictwo PWN, ISBN: 978-83-01-19917-3, 2018

- monografia naukowa stanowiąca antologię najważniejszych technologii analizy sieci złożonych, wraz z autorskimi przykładami i propozycjami rozwiązania wybranych problemów analitycznych[1]
- **80 p** MNiSW

[H2] M. Kamola, *How to Verify Conway's Law for Open Source Projects*, IEEE Access (7) s. 38469–38480, DOI: 10.1109/ACCESS.2019.2905671, 2019

- artykuł demonstrujący metodę i wyniki porównywania sieci zależności modułów oprogramowania i sieci współpracy programistów[2]
- **100 p** MNiSW, JCR **IF=3,745**, WoS=1, Scopus=2, GS=2

[H3] M. Kamola, *Sensitivity of Importance Metrics for Critical Digital Services Graph to Service Operators' Self-Assessment Errors*, Security and Communication Networks, DOI: doi.org/10.1155/2019/7510809, 2019

- artykuł analizujący wrażliwość wybranych indeksów centralności węzłów na błędy oszacowań wagi łącz w sieci[3]
- **40 p** MNiSW, JCR **IF=1,288**, WoS=1, Scopus=1, GS=1

[H4] M. Kamola (70%), P. Arabas, *Network Resilience Analysis: Review of Concepts and a Country-Level Case Study*, Computer Science (15.3) s. 311–327, DOI: 10.7494/csci.2014.15.3.311, 2014

- artykuł analizujący odporność strukturalną sieci polskich systemów autonomicznych[4]
- **12 P** MNiSW, WoS=1, GS=4

[H5] M. Kamola (70%), P. Arabas, *Improving Time-Series Demand Modeling in Hospitality Business by Analytics of Public Event Datasets*, IEEE Access (8) s. 53666-53677, DOI: 10.1109/ACCESS.2020.2980501, 2020

- artykuł demonstrujący model prognozy popytu na usługi hotelarskie, bazujący na treści opisów wydarzeń, z analizą lokalizacji istotnych pojęć w sieci słów języka polskiego[5]
- **100 p** MNiSW, JCR **IF=3,745**, WoS=1, Scopus=1, GS=1

[1] W Wykazie osiągnięć jako pozycja [1.2]
[2] W Wykazie osiągnięć jako pozycja [4.4]
[3] W Wykazie osiągnięć jako pozycja [4.3]
[4] W Wykazie osiągnięć jako pozycja [4.11]
[5] W Wykazie osiągnięć jako pozycja [4.2]

[H6] Mariusz Kamola, *Using Rate Equation for Modeling Triad Dynamics on Instagram*, International Conference on Digital Information Management (ICDIM), IEEE Explore, DOI: 10.1109/ICDIM.2016.7829782, 2016

- artykuł konferencyjny demonstrujący metody zbierania danych i analizy dynamiki struktury sieci społecznej[6]
- **15 P** MNiSW

[H7] M. Kamola, *Analysis of User Story Dynamics on a Medium-Size Social News Site*, International Conference on Computational Collective Intelligence, LNAI 12496, s. 97–109, DOI: 10.1007/978-3-030-63007-2_8, 2020

- rozdział w monografii prezentujący adaptację modelu epidemicznego w celu opisania dynamiki wiadomości w serwisie typu *social news*[7]
- **20 p** MNiSW

[H8] B. Laskowska, M. Kamola (50%), *Grouping compositions based on similarity of music themes*, Plos ONE, DOI: 10.1371/journal.pone.0240443, 2020

- artykuł prezentujący definicję motywu muzycznego oraz towarzyszących mu algorytmów grupowania utworów muzycznych[8]
- **100 p** MNiSW, JCR **IF=2,74**

[H9] M. Kamola (70%), E. Niewiadomska-Szynkiewicz, B. Piech, *Reconstruction of a Social Network Graph from Incomplete Call Detail Records*, Intl. Conference on Computational Aspects of Social Networks (CASoN), IEEE Explore, DOI: 10.1109/CASON.2011.6085932, 2011

- artykuł konferencyjny demonstrujący metodę tworzenia grafu powiązań społecznych z rejestru anonimowych rozmów telefonicznych[9]
- **15 P** MNiSW, WoS=3, Scopus=6, GS=10

[H10] M. Kamola, *Protecting privacy of GPS trails by anonymization of the road graph*, ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics s. 59–62, ACM Digital Library, DOI: 10.1145/2835022.2835033, 2015

- artykuł konferencyjny demonstrujący metodę anonimizacji grafu z zachowaniem istotnych danych na potrzeby dalszej analizy w określonym celu[10]

4.3 Opis osiągnięcia naukowego

Powszechny dostęp do danych masowych stał się faktem. Dzięki obfitości urządzeń pomiarowych i rejestrujących, coraz więcej zjawisk pozostawia cyfrowe ślady. Szacuje się, że wzrost ten ma charakter wykładniczy [1]. Odpowiednie nasycenie przestrzeni śladami cyfrowymi przynosi zmianę jakościową: danych jest tyle, że można je zestawiać i analizować, co prowadzi do powstania nowych informacji, które często wpływają na procesy je tworzące, zamykając w ten sposób pętlę sprzężenia.

Jako społeczeństwo informacyjne, podobnie do naszych pierwotnych przodków doświadczamy osobliwości bezpośredniej konfrontacji. Przed tysiącami lat była to konfrontacja plemienia z innym,

[6] W Wykazie osiągnięć jako pozycja [7.2]
[7] W Wykazie osiągnięć jako pozycja [2.1]
[8] W Wykazie osiągnięć jako pozycja [4.1]
[9] W Wykazie osiągnięć jako pozycja [7.8]
[10] W Wykazie osiągnięć jako pozycja [7.3]

rozumnym plemieniem. Obecnie stajemy czoło w czoło z maszyną realizującą skomplikowany algorytm, nie zawsze rozumiany nawet przez jego twórców. Tamta konfrontacja doprowadziła do powstania cywilizacji [12]; skutki obecnej dopiero poznamy, przy czym wiadomo już, że za przyczyną wykorzystania maszyn będą one kolosalne, bo potęgujące intencje twórców algorytmów.

Jednocześnie jako naukowcy, inżynierowie i twórcy algorytmów, łączymy te cyfrowe ślady, otrzymując bardzo często struktury sieciowe będące naturalnym odzwierciedleniem relacji pomiędzy różnymi obiektami. Interakcje wzajemne użytkowników portali społecznościowych, powiązania w zespole programistów, umowy międzyoperatorskie dostawców Internetu — to przykłady, w których odtworzenie sieci powiązań aktorów jest możliwe i wartościowe. Wynikowe struktury zasadniczo modelują relacje pomiędzy autonomicznymi istotami społecznymi albo wytworami techniki. Okazuje się, że wiele takich sieci (albo grafów, termin równoważny) posiada charakterystyczne cechy, dzięki którym uznawane są za sieci *złożone*. Złożoność sieci rozumiana w ten specyficzny sposób nie wynika wprost z ich rozmiaru, lecz z określonych własności topologicznych. Zarówno w sieciach społecznych, jak i technologicznych, połączenia między węzłami nie mają struktury regularnej, jednocześnie nie będąc zupełnie losowymi. W obrębie tak odlegle postawionych warunków skrajnych, środowisko naukowe nie opracowało jednolitej i precyzyjnej definicji pojęcia sieci złożonej [9] i nie jest odczuwalna presja ani potrzeba jego konkretyzacji. Niemniej istnieje zestaw własności sieci powszechnie utożsamianych z jej złożonością. Bezskalowość, zjawisko małego świata oraz silne lokalne gronowanie to cechy zaobserwowane przez jednego z prekursorów badań, A.-L. Barabásiego [2], i szeroko uznane. Pojęcia te zostaną wkrótce rozwinięte.

4.3.1 Teza i motywacja

Takie właśnie sieci stały się głównym obiektem mojego zainteresowania naukowego w ciągu ostatnich dziesięciu lat. Podchodząc wielokrotnie do analizy danych rzeczywistych w celu uzyskania oryginalnego wyniku badawczego lub celu praktycznego, zauważyłem następujące uwarunkowania prawne, społeczne i techniczne, które stosunkowo szybko zdeterminowały zakres sprawozdawanego osiągnięcia naukowego:

1. Ślady cyfrowe powstają i są gromadzone w olbrzymiej ilości. Podmioty mające możliwości techniczne zbierania danych dążą do ich gromadzenia i przetwarzania, aby realizować własne lub postawione im cele.

2. Podmioty udostępniają zgromadzone dane, jednak te obarczone są szeregiem ograniczeń technicznych i prawnych. Są to ograniczenia ilościowe i jakościowe: zakres, spójność, poprawność, format i dopuszczalny sposób wykorzystania danych. U dominujących dostawców komercyjnych można zaobserwować wyraźną tendencję ograniczania i rosnącej kontroli dostępu do udostępnianych danych. Obserwuję również pewną stagnację w postępie udostępniania danych przez podmioty publiczne.

3. Rekonstrukcja sieci wskutek złączenia danych jest procesem unikatowym, ściśle powiązanym z konkretnym celem badawczym lub inżynierskim. Wykorzystuje ona co prawda standardowe narzędzia i techniki, niemniej dobór ich konkretnego zestawu, dobór parametrów uruchomienia, konieczność dokonywania autorskich zmian i doraźnego konstruowania nowych procedur prowadzą do tworzenia rozwiązań dedykowanych dla poszczególnych problemów.

4. Analiza sieci i dalsze wnioskowanie są pozornie łatwe dzięki istnieniu bogatego instrumentarium — w szczególności w przypadku podejścia wykorzystującego uczenie maszynowe. Dostępność narzędzi umożliwia działania analityczne i modelowanie bez konieczności głębszej znajomości

specyfiki stosowanych technik, a nawet natury modelowanych zjawisk. Może to prowadzić do wyników tyleż spektakularnych, co wrażliwych na niestacjonarność formatów danych i samych opisywanych przez nie procesów. W ostateczności może prowadzić do fałszywych wniosków analiz.

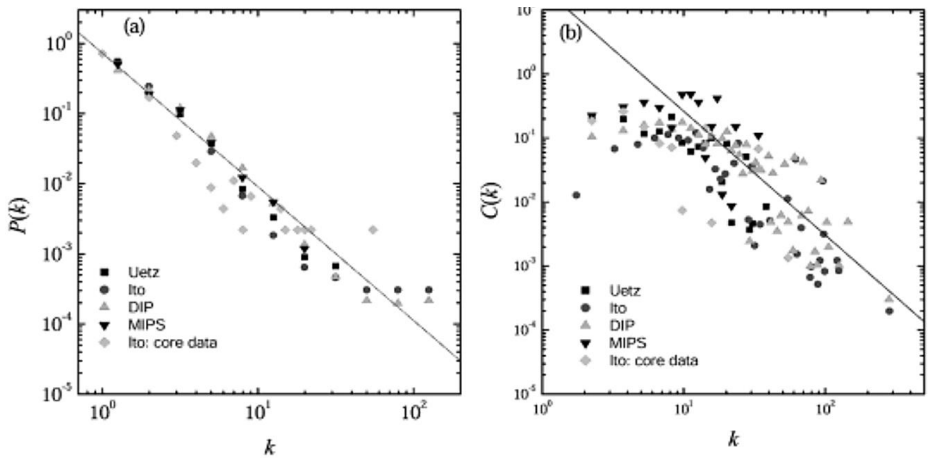
Biorąc pod uwagę powyższe, można postawić tezę, że **prace badawcze, rozwojowe, edukacyjne i popularyzatorskie dotyczące sztuki pozyskiwania danych, łączenia ich w struktury sieciowe, a następnie doboru lub opracowania metod analizy są wartościowe i pożądane, zarówno w kontekście naukowym, jak i społecznym**. Na przedstawiane tutaj osiągnięcie naukowe *Metody analizy i odkrywania struktury sieci społecznych i technologicznych* składają się dokładnie takie prace, opisane cyklem publikacji.

Szczegółowy opis osiągnięcia został podzielony na pięć podrozdziałów. Pierwszy (rozdz. 4.3.2) prezentuje charakterystykę sieci złożonych, których dotyczą metody składające się na osiągnięcie. Kolejny (rozdz. 4.3.3) zawiera omówienie specyfiki osiągnięcia naukowego, z wyszczególnieniem jego elementów składowych. Dwa następne (4.3.4 i 4.3.5) przedstawiają szczegóły metod opracowanych w poszczególnych publikacjach cyklu, w umownym podziale na prace dotyczące analizy i odkrywania sieci złożonych, odpowiednio. Opis kończy podsumowanie w rozdz. 4.3.6.

4.3.2 Sieci złożone — społeczne i technologiczne

Wszystkie publikacje składające się na cykl tematyczny są związane z sieciami złożonymi. Część z nich bezpośrednio wykorzystuje lub bada typowe ich własności. Pozostałe prezentują rozwiązania bardziej uniwersalne, ale w praktyce stosowane głównie w sieciach tego typu.

Z sieciami złożonymi mamy do czynienia zaskakująco często, a jednocześnie ich własności są nieintuicyjne. Dlatego domagają się należytej uwagi, a ich zaniedbanie może prowadzić do błędów w analizie i wnioskowaniu:



Rysunek 1. (a) Rozkład prawdopodobieństwa $P(k)$ dla stopni wierzchołków k w pięciu różnych sieciach oddziaływań białek, w skali podwójnie logarytmicznej, z aproksymacją linią prostą. (b) Rozkład wartości współczynnika gronowania $C(k)$ w zależności od stopnia wierzchołka k (źródło: [36], Fig. 2).

- *Bezskalowość.* Prawdopodobieństwo $P(k)$ wystąpienia w sieci węzła o stopniu, tj. liczbie powiązań, równym k ma rozkład podobny do rozkładu potęgowego: $P(k) \sim k^{-\alpha}$, por. rys. 1a. Jedną z konsekwencji jest to, że końcowy fragment (tzw. ogon) rozkładu jest podobny do całości. Nie można więc wyróżnić żadnego charakterystycznego zakresu k, a więc i skali wykresu rozkładu stopni, w którym zaobserwowalibyśmy zdecydowany zanik prawdopodobieństwa wystąpienia węzłów o wielkich stopniach. Prawdopodobieństwo powstania w sieci wielkiego węzła jest znaczne, tak jak znaczne jest prawdopodobieństwo kataklizmów przyrodniczych, np. tsunami. Ten fakt wymusza odstąpienie od tradycyjnych, statystycznych metod analizy, gdyż w pewnych przypadkach rozkład stopni wierzchołków może mieć nieskończoną wariancję, a nawet — wartość oczekiwaną.

- *Zjawisko małych światów.* Średnia odległość między węzłami sieci jest zaskakująco mała, to znaczy rośnie bardzo powoli, co najwyżej logarytmicznie wraz ze wzrostem sieci. Może to wynikać z bezskalowości sieci, ale nie musi — co zostało poparte modelami matematycznymi. Wiedza o tym zjawisku została skutecznie spopularyzowana, ale z reguły nie znajduje ona odbicia w praktycznych zastosowaniach uwzględniających szybkość propagacji informacji w sieci, np. skutecznemu zapobieganiu pandemii.

- *Silne lokalne gronowanie.* W wielu sieciach, głównie reprezentujących relacje społeczne, występują liczne, nieduże ale mocno wewnętrznie połączone grupy węzłów o niewysokich stopniach. W ujęciu statystycznym odpowiada temu malejąca wartość współczynnika gronowania $C(k)$ wraz ze wzrostem stopnia węzła k, por. rys. 1b. $C(k)$ jest obliczane jako iloraz zaobserwowanych i wszystkich możliwych połączeń pomiędzy sąsiadami wierzchołków o założonym stopniu k. Zjawisko silnego gronowania jest znane specjalistom i zostało wykorzystane w rozbudowanych modelach matematycznych dyfuzji informacji. Jednak prace te w zaskakująco małym stopniu przekładają się na działanie praktyczne, zwłaszcza dotyczące jego niepożądanych skutków społecznych, np. powstawania baniek informacyjnych (*echo chambers*). Świadomość niejednorodności sieci jest kluczowa dla modelowania zachodzących w niej zjawisk dynamicznych, w odniesieniu tak do stanu węzłów, jaki i samej struktury sieci.

Co ciekawe, powyższe własności dają się zauważyć tak w sieciach społecznych, jak i technologicznych. Przez sieci społeczne rozumiemy struktury ukształtowane niejako oddolnie i w wyniku autonomicznych decyzji węzłów-jednostek społecznych (niekoniecznie ludzkich). Natomiast sieci technologiczne są z reguły wytworem bardziej scentralizowanego procesu planistycznego. Można do nich zaliczyć np. sieć wodociągową, połączeń lotniczych, układ elektroniczny. Nie jest to klasyfikacja ścisła lecz, podobnie jak sama definicja sieci złożonej, wskazująca jedynie bardzo reprezentatywne przykłady. Natomiast w obrębie tych przykładów zaobserwowano kolejne własności, które z czasem zaczęto wtórnie traktować jako wyznaczniki sieci obu typów. Taką własnością jest *asortatywność*, czyli tendencja do tworzenia się powiązań w sieci pomiędzy węzłami o podobnych stopniach. Typowo objawia się ona w sieciach powiązań międzyludzkich. Natomiast *dysasortatywność*, cecha odwrotna, charakteryzuje sieci technologiczne. Inne popularne metody oceny własności sieci bazują na zliczaniu w niej charakterystycznych struktur, np. trójkątów.

4.3.3 Elementy składowe osiągnięcia

Specyfika postawionej tutaj tezy o istotności głębokiego rozumienia istoty sieci złożonych i celowości oryginalnych adaptacji lub wręcz tworzenia dedykowanych algorytmów ma swoje odbicie w strukturze cyklu publikacji. Zasadniczo prezentuje on zbiór wytworzonych metod i algorytmów, powiązanych horyzontalnie i demonstrujących zasadność tezy w określonym obszarze zastosowań

praktycznych. Publikacje odnoszą się do siebie głównie poprzez wykorzystywanie podobnego podejścia projektowego, osiąganie podobnych celów lub korzystanie z podobnych analogii. Cykl publikacji jest więc przejawem eksplorowania możliwie wielu aspektów analizy sieci złożonych, zamiast doskonalenia pojedynczej metody badawczej. Jednocześnie opracowane metody z reguły stanowią oryginalny wkład badawczy poszerzający aktualny stan wiedzy.

Sformułowanie postawionej tutaj tezy nie nastąpiło a priori lecz wyłoniło się wskutek początkowo odrębnych badań prowadzonych w celach szczegółowych. Jej pełnym wyrazem jest otwierająca cykl monografia *Sieci społeczne i technologiczne. Jak zrozumieć, jak wykorzystać* [H1] autorstwa mojego i dra Piotra Arabasa. Wraz ze współautorem postawiliśmy sobie za cel ukazanie istotności cech sieci złożonych w sposób sugestywny, atrakcyjny. Dlatego dużą wagę przyłożyliśmy do opracowania danych reprezentujących sieci rzeczywiste i odnoszące się do polskich realiów, a przez to bliższe czytelnikom. Sieci te wykorzystywane są do rozwiązywania zadań praktycznych zarówno za pomocą technik klasycznych, jak i nowych algorytmów zaproponowanych przeze mnie w tym celu.

Monografia nie jest zatem sumą wcześniejszych moich prac badawczych, lecz ich ogólniejszą konsekwencją, tj. manifestacją poglądu o konieczności głębokiego rozumienia przez analityka specyfiki sieci, a także o sensowności konstruowania własnych algorytmów analitycznych. Aby ułatwić rozumienie, prezentujemy uznane wskaźniki liczbowe oraz struktury opisujące własności całej sieci (asortatywność, współczynniki gronowania, rdzenie, kliki, triady itp.), jak również wskaźniki opisujące poszczególne węzły (różne rodzaje centralności). Pojęcia te odnoszą się do sieci w ujęciu statycznym, gdyż aparat matematyczny służący analizie dynamiki sieci nie okrzepł jeszcze do formy kanonu metod. Na przedstawiony przeze mnie wybór istniejących metod analizy zmian w sieci składają się prace klasyczne, np. eksperymenty Milgrama [27] i Harary'ego [17], badające rozwój kaskad i stabilność strukturalną sieci, odpowiednio. Natomiast ewolucja stanu węzłów opisywana modelami liniowymi umożliwia szerokie zastosowanie metod analizy układów dynamicznych i teorii sterowania, i w takim ujęciu została ona omówiona. Co ciekawe, wątek badań strukturalnej sterowalności sieci złożonych jest nadal kontynuowany przez zespół Barabásiego [25].

Aby przekonać o sensowności konstruowania własnych algorytmów, przedstawiam w monografii dwie autorskie metody analizy danych. Walor tworzenia własnego podejścia polega na tym, że może ono uwzględniać specyficzną wiedzę dziedzinową twórców — np. w [H8] parametry algorytmu grupowania utworów muzycznych mogły i powinny zostać dobrane ekspercko. Dwie wspomniane metody z kolei inspirowane są pojęciami z teorii wsparcia decyzyjnego. Traktuję je jako osiągnięcia naukowe, a przez to i całą monografię jako element cyklu publikacji. Podobnie jak kilka innych osiągnięć, nie zasadzają się one ściśle na własnościach sieci złożonych, mając bardziej uniwersalny charakter. Jednak ich obszar zastosowań dotyczy głównie sieci złożonych.

Monografię uzupełnia praktyczny instruktaż posługiwania się oprogramowaniem służącym pozyskiwaniu, składowaniu, analizie i prezentacji danych tak, aby dać czytelnikowi wsparcie w całym zakresie jego działań: od zrozumienia cech szczególnych sieci do rzeczywistego wykorzystania dostępnych danych. Dobór prezentowanych zagadnień praktycznych jest skutkiem eksperckiego spojrzenia obu autorów na dziedzinę i własnych wieloletnich doświadczeń.

Publikacje powstałe po monografii [H1] są wynikiem badań prowadzonych w tym samym duchu i składają się wraz z wcześniejszymi pracami na przekrojowy zbiór metod, argumentujący tezę.

Tytuł osiągnięcia — *Metody analizy i odkrywania struktury sieci społecznych i technologicznych* — odzwierciedla jego dwuczęściową strukturę. Poprzez metody analizy rozumiem dokonania umożliwiające ocenę istniejących sieci przy założeniu, że są one możliwie wiernym modelem zjawisk rzeczywistych. Poprzez odkrywanie struktury rozumiem sam proces tworzenia takiego sieciowego modelu — wychodząc od danych surowych, albo na podstawie innej sieci, albo w sposób mieszany.

W części dotyczącej *analizy* sieci wyróżniają się trzy następujące wątki badawcze:

(A) Pierwszy dotyczy wzbogacenia wiedzy o sieci poprzez wskazanie nowych sposobów korzystania ze standardowych i uznanych metod analizy sieci, a także wprowadzenia nowych wskaźników. W publikacjach [H1] i [H2] proponuję nowe metody oceny jakości grupowania aglomeracyjnego węzłów sieci, uwzględniające fakt, że są to sieci powstałe z sieci dwudzielnych. Natomiast w [H3] i [H4] badam wrażliwość indeksów centralności na zakłócenia informacyjne oraz proponuję nowe indeksy niezawodności, odpowiednio.

(B) Kolejny wątek analityczny obejmuje badanie zależności pomiędzy strukturą sieciową a danymi zewnętrznymi w stosunku do niej. W publikacji [H5] siecią tą jest sieć słów języka, a danymi zewnętrznymi — popyt na usługi hotelarskie w dni imprez masowych opisanych tymi słowami. Natomiast w [H2] mamy do czynienia z dwiema sieciami i propozycją zdefiniowania optymalnego odwzorowania pomiędzy nimi.

(C) Trzeci wątek analityczny dotyczy zjawisk dynamicznych: powstawania nowych połączeń w sieci [H6] i zmiany stanu jej węzłów [H7]. W obu przypadkach do opisu tych zjawisk adaptuję z powodzeniem parametryczne modele fizykalne znane z innych nauk.

W części dotyczącej *odkrywania struktury* sieci zupełnie nowej albo wynikającej z już istniejącej zaznaczają się również trzy wątki badawcze:

(D) Pierwszy obejmuje metody identyfikacji faktu istnienia połączenia pomiędzy węzłami nowotworzonej sieci lub określenia siły tego powiązania. Łączy je stosowanie podejścia eksperckiego, czerpiącego z wiedzy w konkretnych dziedzinach. We wspomnianej już pozycji [H2] dziedziną tą jest inżynieria oprogramowania. Publikacja [H8] dotyczy teorii muzyki. Natomiast prezentowana druga przykładowa metodyka w monografii [H1] czerpie z algorytmów przetwarzania języka naturalnego.

(E) W drugim wątku stosuję podejście komplementarne, dobierając działanie metod tak, by uzyskać sieć o założonych cechach. Wynika to z obecnej wiedzy, że sieci określonej proweniencji są do siebie podobne i tworzą rodziny [28], np. sieci sąsiadowania słów w językach czy interakcji białkowych — niezależnie od konkretnego języka i organizmu. W publikacji [H9] rekonstruuję sieć powiązań między klientami operatora telefonii, zakładając, że powinna być ona bezskalowa. Z kolei w pozycji [H6], wymienionej już w kontekście badania dynamiki, struktura sieci jest odkrywana tak, aby wydobyć efekt silnego lokalnego gronowania.

(F) Ostatni wątek demonstruje istotny wpływ aspektu prywatności informacji na proces opracowywania algorytmów. Obejmuje on dwa przypadki komplementarne. Wspomniana już rekonstrukcja sieci w [H9] jest dokonywana dla danych wejściowych ze zanonimizowanymi identyfikatorami abonentów i numerów telefonicznych. Zaproponowana metoda rekonstrukcji powiązań między abonentami jest próbą wydobycia możliwie najpełniejszej informacji z tak spreparowanych danych wejściowych. Natomiast w [H10] wykonuję zadanie odwrotne, opracowując metodę usuwania wybranych informacji z danych tak, aby spreparowana sieć wciąż dostarczała użytecznych informacji, jednakże w wybranym, kontrolowanym zakresie.

Tabela 1. Zestawienie wątków poruszanych w sposób istotny w poszczególnych publikacjach.

		H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
analiza	A	•	•	•	•						
	B		•			•					
	C						•	•			
odkrywanie	D	•	•						•		
	E						•			•	
	F									•	•
Big Data					•		•	•		•	•
NLP		•				•			•		

Publikacje te wiążą się ze sobą poprzez poruszanie tych samych wątków, poprzez wspólne cele, metody badawcze i klasy źródeł danych. Zestawienia wątków obecnych w publikacjach dokonano w tabeli 1. Ponadto, duża część osiągnięć wymagała użycia specyficznych technik analizy danych wejściowych o znacznym wolumenie i różnorodności (Big Data) lub analizy języka naturalnego (NLP, *natural language programming*). Te cechy wspólne podejść zostały uwzględnione w dwu końcowych wierszach zestawienia.

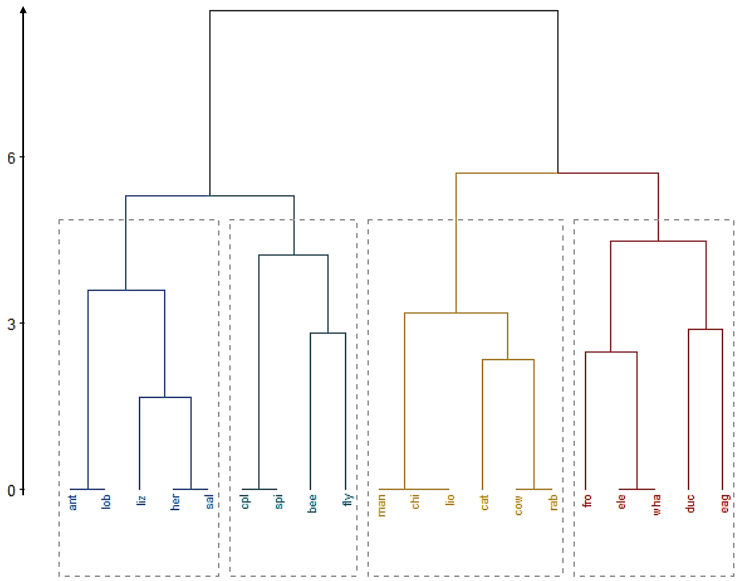
Poniższe dwa rozdziały zawierają omówienie poszczególnych publikacji. Prezentację każdej z nich rozpoczyna śródtytuł. Poszczególne osiągnięcia szczegółowe zostały sformułowane hasłowo i wyróżnione czcionką pogrubioną, ze wskazaniem na marginesie identyfikatora wątku badawczego, którego dotyczą.

4.3.4 Analiza własności sieci

Rozwiązując bardziej złożone zadanie analityczne można albo stosować kombinacje gotowych metod, albo tworzyć metody nowe, dedykowane dla konkretnego zadania. Dostępność metod gotowych, oferowanych w postaci pakietów obliczeniowych skutkuje łatwością ich stosowania bez konieczności zaznajomienia się z istotą działania, co jest praktyką niezdrową lecz powszechną. W moich badaniach odniosłem się do obu scenariuszy, proponując laikom metody oceny jakości działania kombinacji gotowych algorytmów oraz proponując zupełnie nowe algorytmy analityczne.

Metodyka dwukryterialnej oceny jakości grupowania aglomeracyjnego w sieci dwudzielnej

Sieci dwudzielne, i ogólnie sieci z wydzielonymi typami węzłów, stanowią ważną reprezentację danych, umożliwiając przedstawienie różnych typów relacji między obiektami. Struktury takie są chętnie wykorzystywane w zadaniach ekstrakcji informacji, mechanizmach rekomendacji czy analizie języka naturalnego. W sieci dwudzielnej, tj. zawierającej węzły typu A oraz B, krawędzie z A do B służą modelowaniu atrybutów (B) posiadanych przez węzły (A) lub działań zachodzących pomiędzy węzłami różnych typów. Wspólnym problemem w analizie takich sieci jest należyta projekcja, czyli rekonstrukcja powiązań pomiędzy węzłami typu A na podstawie informacji o ich powiązaniach z sąsiadami typu B. Sieci dwudzielne może cechować cały zakres gęstości powiązań pomiędzy węzłami obu typów. Dobór właściwej metody projekcji zależy od wyjściowej gęstości sieci dwudzielnej, lecz przede wszystkim powinien odpowiadać sensowi zjawiska modelowanego przez sieć. Dla przykładu, z sieci reprezentującej przebywanie osobników (węzły typu A) w rozmaitych miejscach (węzły typu B) inaczej rekonstruujemy sieć propagacji epidemii, a inaczej — sieć intensywności relacji społecznych.

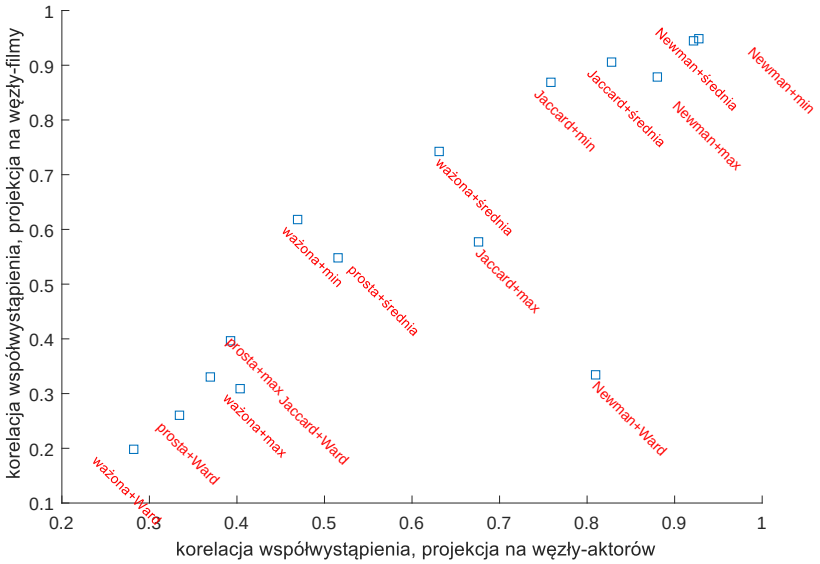


Rysunek 2. Przykładowy dendrogram. Segmenty poziome odpowiadają operacjom złączenia pojedynczych węzłów lub ich grup. Rzędne segmentów oznaczają odległość pomiędzy grupami w momencie ich łączenia. Jest to odległość współwystąpienia (*cophenetic distance*) dla par węzłów przynależnych dotąd do różnych grup. Kolejność kroków grupowania odpowiada rosnącym rzędnym segmentów poziomych, rosnąco (źródło: Wikipedia).

Innym powszechnym problemem w analizie sieci jest wyodrębnienie stosunkowo luźno połączonych między sobą podsieci, lecz o dobrze rozwiniętej strukturze połączeń wewnętrznych. Przeprowadza się je zarówno dla sieci z węzłami naturalnie jednorodnymi, jak dla sieci rekonstruowanych z sieci dwudzielnych. Powszechnie wykorzystuje się w tym celu algorytmy grupowania aglomeracyjnego, łączące w kolejnych krokach najbliższe sobie węzły lub grupy węzłów. Powstaje dendrogram — zapis procesu grupowania, na podstawie którego dokonuje się ostatecznego podziału sieci, por. rys. 2. Dobór metody obliczania odległości łączonych grup determinuje strukturę dendrogramu i powinien odpowiadać naturze zadania. Użytkownicy pakietów obliczeniowych najczęściej wybierają spośród czterech metod obliczania odległości: minimum, maksimum, średniej lub Warda. Każda agreguje odległości pomiędzy poszczególnymi parami punktów adekwatnie do swojej nazwy (metoda Warda oblicza przyrost wariancji wskutek połączenia grup). Umiejętności techniczne i doświadczenie wymagane od użytkownika w celu skorzystania z grupowania są nieporównywalnie mniejsze od potrzebnych umiejętności analitycznych umożliwiających wybór właściwego wariantu. Jeśli zadanie to jest częścią większego problemu, którego inne etapy rozszerzają przestrzeń parametrów decyzyjnych, liczba i sens możliwych kombinacji utrudniają nieprofesjonalistom systematyczną ocenę uzyskanych wyników i wybór dobrego połączenia metod.

A W [H1], s. 120 proponuję **metodykę oceny jakości grupowania aglomeracyjnego węzłów w sieci dwudzielnej**. Wyniki grupowania zależą od wyboru metody projekcji w sieci dwudzielnej oraz od wariantu obliczania odległości między grupami. Rozważam cztery standardowe rodzaje projekcji: prostą, ważoną, Newmana i Jaccarda. W połączeniu z czterema wariantami obliczania odległości daje to 16 kombinacji. Wskazuję przestrzeń dwóch kryteriów oceny jakości grupowania: kryteriami tymi są wartości korelacji współwystąpienia (*cophenetic correlation*, [H1] s. 117), czyli korelacji pomiędzy faktycznymi odległościami par węzłów w sieci a odpowiadającym im odległościami współwystąpienia. Obie korelacje obliczane są dla rekonstruowanych sieci: węzłów typu A

oraz węzłów typu B (czyli dla projekcji w odwrotnym kierunku). Otrzymane wartości współczynników współwystąpienia przedstawiam na wykresie w tej przestrzeni, co umożliwia ich wygodne porównywanie i wskazanie rozwiązań optymalnych w sensie Pareto. Rys. 3 prezentuje przykładowe wyniki dla sieci dwudzielnej polskich aktorów i filmów z ich udziałem.

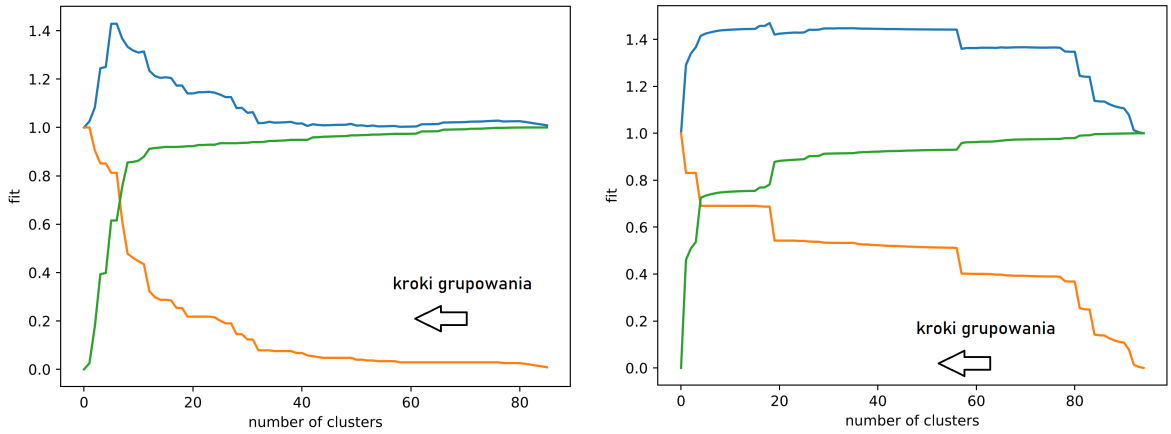


Rysunek 3. Współczynniki współwystąpienia dla projekcji sieci aktorów i filmów (źródło: [H1], rys. 6.10).

W zadaniach analitycznych zazwyczaj wykonuje się projekcję tylko w jednym kierunku, np. rekonstruując powiązania między dokumentami (A) na podstawie zawartych w nich pojęć (B). Korelacja współwystąpienia dla grupowania jest tym wyższa, im bardziej odległość między grupami łączonymi w kroku algorytmu aglomeracyjnego odpowiada rzeczywistej odległości w sieci pomiędzy poszczególnymi węzłami łączonych grup. Mój postulat rozważenia konsekwencji wyboru algorytmu dla zrekonstruowanej sieci drugiego typu (tu: pojęć) zasadza się na dualizmie sieci dwudzielnych. Struktura takich sieci implikuje jednocześnie relacje między węzłami A i B, toteż i metody rekonstrukcji tych relacji powinny uwzględniać jakość *obu* rekonstruowanych sieci. Jeżeli nasz wybór algorytmu prowadzi jednocześnie do bardzo dobrego pogrupowania dokumentów, skutkując fatalnym pogrupowaniem pojęć, to taka dysharmonia powinna skłonić do głębszego rozważenia istoty naszego zadania. Albo do wyboru innego, równie dobrego algorytmu. W zbiorze rozwiązań przedstawionym na rys. 3 widać dużą zależność pomiędzy wartościami obu kryteriów, lecz zdarzają się też odstępstwa (por. położenie punktów “Newman+Ward” i “ważona+min”).

Procedura porównywania zgodności wyników grupowania w sieci dwudzielnej

Opracowane przeze mnie algorytmy analizy sieci złożonych w większości przypadków mają zastosowanie praktyczne, jak w powyższym osiągnięciu, oraz stanowią rozwinięcie lub polemikę z istniejącymi podejściami. Charakter polemiczny mają moje badania tzw. prawa Conwaya [10]. Zauważył on, że struktura wytwarzanego systemu informatycznego odzwierciedla strukturę organizacyjną jego wytwórcy. Wynika to z ergonomii pracy, a zwłaszcza z jej nakładu na wprowadzenie zmian w istniejącym oprogramowaniu. Jeśli zmiany te wymagają współudziału wielu jednostek, narzut koordynacyjny wprowadza koszty, ryzyka i opóźnienia. Dlatego granice funkcjonalności modułów oprogramowania nie bez powodu przebiegają po granicach kompetencji i odpowiedzialności współpracujących zespołów programistycznych.

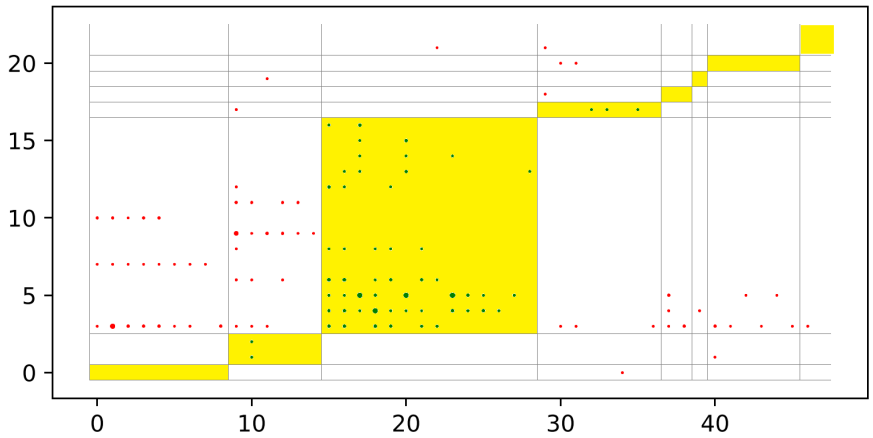


a) b)

Rysunek 4. Wartości J_{TP} (pomarańcz.) i J_{TN} (ziel.) unormowanych do zakresu $\langle 0; 1\rangle$ oraz ich sumy (nieb.) dla projektu Keras. Wykres sumy osiąga maksimum dla odmiennej liczby grup w przypadku grupowania modułów (a) i w przypadku programistów (b). Metodyka porównywania sieci uwzględnia jej podział we wszystkich finalnych krokach grupowania, począwszy od tego, w którym wystąpiło maksimum (źródło: [H2], Fig. 3).

Podjąłem próbę weryfikacji tej prawidłowości dla zbioru dużych i udanych projektów o otwartym kodzie źródłowym. (Użyte metody odkrywania sieci programistów i sieci modułów oprogramowania zostały, zgodnie ze swoim charakterem, zaprezentowane w rodz. 4.3.5.) Pierwszym z osiągnięć analitycznych przedstawionych w [H2], stanowiącym podstawę dalszych prac, jest **algorytm wyznaczania optymalnej liczby grup w procesie grupowania aglomeracyjnego** (modułów albo programistów). Wprowadzam dwie funkcje monotoniczne względem kolejnych kroków grupowania aglomeracyjnego. Pierwsza, $J_{\mathrm{TP}}(G, k)$, sumuje odległości par węzłów z sieci G, które znalazły się w tej samej grupie w kroku k grupowania aglomeracyjnego. Druga, $J_{\mathrm{TN}}(G, k)$, sumuje dopełnienia odległości par węzłów z sieci G, które znalazły się w różnych grupach w kroku k (dopełnienie do średnicy sieci G). Symbole i charakter tych funkcji odpowiadają kategoriom klasyfikacji w macierzy pomyłek: TP (*true positive*) określa profil kosztów grupowania modułów, podczas gdy TN (*true negative*) określa profil kosztów rozdzielenia modułów. Suma J_{TP} i J_{TN}, unormowanych uprzednio do przedziału $\langle 0; 1\rangle$, oznacza łączny koszt grupowania. Taką analizę grupowania prowadzę względem obu sieci: modułów i programistów. Przykładowe wykresy przedstawia rys. 4. Krok, w którym osiągane jest maksimum łącznego kosztu grupowania, określa wyjściowy podział sieci. Wyznaczona optymalna liczba grup z reguły jest różna dla sieci modułów (ozn. $N^{\star}_{M}$) i sieci programistów (ozn. $N^{\star}_{D}$).

Drugim elementem składowym procedury i osiągnięciem analitycznym jest **zdefiniowanie stopnia podobieństwa strukturalnego dwóch sieci** w kontekście wyników grupowania aglomeracyjnego. Dla rozpatrywanego zagadnienia wartość tego stopnia jest miarą prawdziwości stwierdzenia "struktura A odpowiada strukturze B" w prawie Conwaya. Z uwagi na złożoność obliczeniową, upraszczam problem do zadania optymalnego utożsamienia N grup programistów z N grupami modułów, gdzie $N \leq \min(N^{\star}_{M}, N^{\star}_{D}, 10)$. Sprawdzam jakość każdego z $N!$ możliwych odwzorowań, którą definiuję jako bilans sumy aktywności programistów dla modułów w obrębie grup utożsamionych ze sobą i podobnej sumy dla grup nieutożsamionych. Oba składniki tej miary są rozwinięciem koncepcji funkcji J_{TP} i J_{TN}. Jako ostateczne zostaje wybrane odwzorowanie dające najlepszy bilans.



Rysunek 5. Przykładowe utożsamienie grup modułów kodu (kolumny macierzy) z grupami programistów (wiersze macierzy) dla projektu TensorFlow. Utożsamione grupy oznaczono kolorem żółtym (źródło: [H2], Fig. 9b).

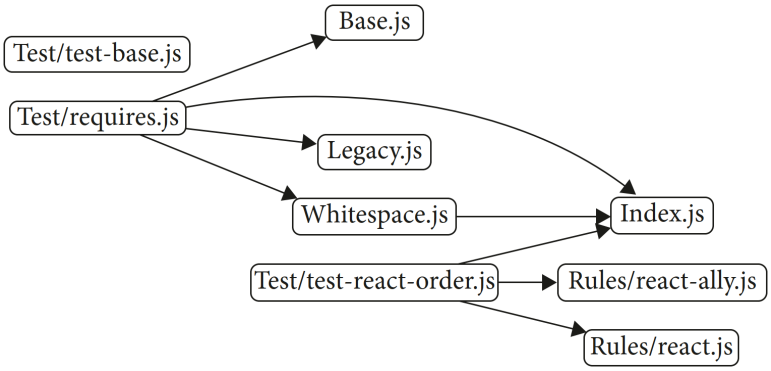
Jeżeli przedstawimy utożsamione grupy jako klatki na antyprzekątnej macierzy aktywności programistów, wówczas składniki miary odpowiadają sumom elementów w tych klatkach oraz sumom poza nimi — por. rys. 5. Poszukiwanie optymalnego odwzorowania jest NP-trudne, dlatego ograniczam liczbę porównywanych grup, korzystając w razie potrzeby z wyników z dalszych faz grupowania (tj. operując na zmniejszającej się liczbie grup).

Metodyka oceny wrażliwości centralności węzłów na błędy szacowania wag połączeń

Inne ważne zagadnienie analityczne dotyczy trafnego, z perspektywy określonego zastosowania, doboru wykorzystywanego indeksu centralności węzłów w sieci. Powszechnie używane są w tym celu indeksy bardzo proste, np. stopień wierzchołka, indeksy nieco bardziej złożone, jak pośrednictwo, oraz zaawansowane wskaźniki widmowe. Moje prace w tym zakresie były prowadzone w kontekście projektu NPC,[11] por. rozdz. 5.4, i dotyczyły sieci uzależnionych od siebie usług kluczowych i cyfrowych (np. dostawa energii, platformy handlowe, usługi bankowe). Specyfika takiej sieci polega na tym, że o ile sam fakt powiązania usług, a więc struktura sieci, jest uznawany za pewny, to siła tego powiązania jest jedynie szacowana przez usługobiorcę. Jej wartość, tożsama z wagą połączenia, jest subiektywna, i należy uznać ją za obarczoną błędem. Zadanie badawcze polega zatem na ocenie wrażliwości dostępnych wskaźników centralności węzłów-usług na tego rodzaju błędy, w celu wskazania najodpowiedniejszych.

Istotną przeszkodę praktyczną w realizacji zadania stanowił brak wiarygodnych danych, tj. przykładowych sieci usług, jak również brak przesłanek co do sposobu wygenerowania sieci testowej pod nieobecność sieci rzeczywistej. Zaproponowałem przyjęcie założenia, że bliską analogią sieci usług kluczowych jest sieć zależności modułów w programie komputerowym. Analogia wynika stąd, że sam fakt zależności jest łatwo identyfikowalny, np. poprzez statyczną analizę kodu. Niekiedy analiza taka upraszcza się do weryfikacji listy modułów włączanych do kodu źródłowego. I dalej, analogicznie, o ile łatwo jest takie związki wykryć, o tyle trudno jest określić ich siłę. Takie spostrzeżenie utorowało drogę do dalszych badań dla rzeczywistych sieci zależności modułów oprogramowania, odtworzonych z dostępnych publicznie w serwisie github.com repozytoriów projektów. To samo źródło danych zostało wykorzystane w [H2]. Rys. 6 przedstawia część faktycznej sieci zależności modułów (niespójność i cykle są zjawiskami normalnym w obu typach sieci). Przez **A** oznaczmy macierz sąsiedztwa sieci.

[11] Projekt nr CYBERSECIDENT/369195/I/NCBR/2017 finansowany przez NCBiR, nazwa: *Narodowa Platforma Cyberbezpieczeństwa* (NPC) — w Wykazie osiągnięć jako pozycja [9.1]



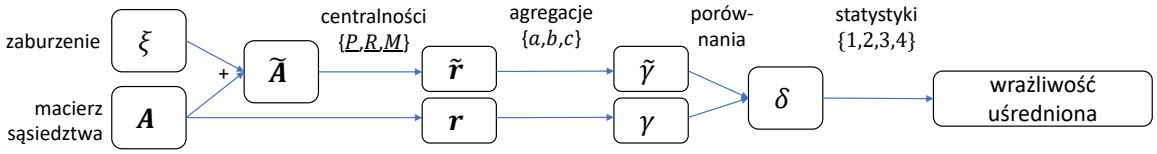
Rysunek 6. Fragment sieci zależności modułów dla projektu Airbnb (źródło: [H3], Fig. 2).

A W [H3] przedstawiam **metodykę oceny wrażliwości wybranych indeksów istotności węzłów** na błędy w oszacowaniu wag łączy, interpretowanych jako siła powiązań usług reprezentowanych przez węzły. Wagi mogą przyjmować wartości całkowite od 1 do 10. W badaniu zostały wzięte pod uwagę trzy indeksy istotności, określające podatność (w znaczeniu strukturalnym, nie programistycznym) węzła-usługi v_k na awarie innych węzłów-usług w sieci:

- Page Rank [32], agregujący rekurencyjnie istotność wszystkich usług wpływających na v_k,
- Reach Centrality — odsetek wszystkich usług, których awaria może dotknąć v_k,
- Maximum Input — centralność obliczana rekurencyjnie, podobnie do Page Rank, ale z uwzględnieniem tylko usług-sąsiadów o największym wpływie.

Oznaczmy przez **r** wektor istotności węzłów.

Zaproponowałem, aby w celu zagregowania istotności poszczególnych węzłów do pojedynczego wskaźnika skalarnego γ dla całej sieci użyć: (a) wartości średniej, (b) mediany, (c) wartości maksymalnej istotności. W celu oszacowania wrażliwości zaproponowałem zaburzenie ustalonej wartości początkowej siły powiązania rozkładem dyskretnym jednostajnym ξ. Zmienna losowa modeluje pomyłkę podmiotu świadczącego usługę v_k w samoocenie siły powiązania tej usługi z usługą ją wspierającą. Pomyłka ta może wynosić maksymalnie N stopni w górę i w dół skali, gdzie N jest parametrem modelu. Błąd δ oceny wskaźnika γ jest obliczany jako odchyłka od wartości rzeczywistej, bazującej na sieci o niezaburzonych wagach, por. rys. 7. Alternatywnie, proponuję obliczanie błędu jako rozbieżności rankingów najistotniejszych węzłów w obu wersjach sieci. Agregacja ww. odchyłek δ dla różnych instancji **A**, czyli różnych sieci, również może być przeprowadzona w wariantach bazujących na średniej lub odchyleniu standardowym — w odniesieniu wartości względnych lub bezwzględnych (cztery kombinacje, ozn. 1–4).



Rysunek 7. Fazy obliczeń wrażliwości uśrednionej oraz wyniki pośrednie. Strzałki oznaczają obliczenia wg różnych wariantów.

Interpretacja uzyskanych wyników w przestrzeni zaproponowanych opcji $\{\underline{\mathrm{P}}, \underline{\mathrm{R}}, \underline{\mathrm{M}}\} \times \{\mathrm{a}, \mathrm{b}, \mathrm{c}\} \times \{1, 2, 3, 4\}$ jest trudna. Wprowadziłem *kryteria wartościujące wyniki*, uznając za lepsze kombinacje dające:

1. małą wrażliwość średnią, niezależnie od projektu i wielkości zaburzenia N;
2. dużą wrażliwość największego obserwowanego błędu, max $|\mathbf{r}-\widetilde{\mathbf{r}}|$, w poszczególnych sieciach na zmianę N;
3. małe odchylenie standardowe największych błędów dla poszczególnych sieci.

Przedstawienie tej trójkryterialnej przestrzeni ocen w trzech rzutach płaskich umożliwia graficzną ocenę rozwiązań, a także wyznaczenie zbioru rozwiązań Pareto-optymalnych. Sensem tego osiągnięcia, podobnie jak metodyki oceny grupowania aglomeracyjnego [H1](A), jest zdefiniowanie przestrzeni kryteriów oceny jakości funkcjonowania wielu kombinacji standardowych metod numerycznych (wyznaczania centralności wierzchołka, agregacji błędów w obrębie pojedynczej sieci, a także w obrębie zbioru rozpatrywanych sieci).

Badania wymagały pobrania kodu źródłowego projektów. Wytypowano i przeanalizowano repozytoria pięciu projektów dużych i popularnych aplikacji lub bibliotek. Do podstawowej analizy zależności modułów wykorzystano oprogramowanie Madge. Zasadniczej analizy dokonano za pomocą autorskiego oprogramowania w języku python z wykorzystaniem bibliotek obliczeń numerycznych.

Procedura oceny odporności strukturalnej sieci systemów autonomicznych

Pozostając przy tematyce badania odporności sieci na zaburzenia, należy przywołać również klasyczne pojęcie odporności sieci infrastrukturalnej, czyli odporności na awarie łączy i węzłów w sieci o jednakowych wagach powiązań. W moich pracach nie zaniedbuję tego wątku; w [H4] badamy wraz ze współautorem odporność polskiego fragmentu sieci Internet na awarie. Z uwagi na szczególną, warstwową naturę tej sieci już sama definicja takiej struktury danych staje się wyzwaniem. Mając świadomość coraz powszechniejszej wirtualizacji usług powodującej powstawanie de facto nowych warstw komunikacyjnych (*overlay networks*), staraliśmy się zachować podejście pragmatyczne, tj. wykorzystywać takie źródła danych, aby otrzymać sieci kompletne i jednocześnie użyteczne, czyli o strukturze zgodnej z realiami biznesowymi operatorów.

Model sieci bazujący na grafie połączeń systemów autonomicznych (*autonomous system*, AS) w zupełności odpowiada przyjętym założeniom. Tablice trasowania protokołu międzyoperatorskiego (*border gateway protocol*, BGP) odzwierciedlają autentyczne założenia biznesowe operatorów. Użycie tablic BGP umożliwiło odtworzenie sieci połączeń prawie 2000 systemów autonomicznych.

W [H4] wprowadzam **metodę oceny podatności węzłów-systemów autonomicznych na awarie** łączy międzyoperatorskich, jak i samych AS. Wykorzystuje ona w oryginalny i specyficzny sposób graf skierowany $G(V,E)$ ze zbiorem $V=\{v_i\}$ wierzchołków-systemów i zbiorem $E=\{e_{ij}\}$ krawędzi-umów międzyoperatorskich. Krawędź e_{ij} reprezentuje świadczenie przez AS v_j usługi dostępu do Internetu dla v_i. Definiuję istnienie trasy BGP między określoną parą $\{v_i,v_j\}$, jeśli istnieje co najmniej jeden system autonomiczny osiągalny jednocześnie z v_i i z v_j. Oznaczam przez $R_{G(V,E)}(v_i)$ zbiór tych AS, do których istnieją trasy BGP z v_i. Proponuję definicję wskaźnika podatności strukturalnej v_i na awarię uzależnionego od łącznej liczby utraconych połączeń do pozostałych AS z tytułu awarii pojedynczych łączy: (A)

$$u_{\mathrm{E}}(v_i)=\frac{\sum_{e\in E}||R_{G(V,E))}(v_i)\backslash R_{G(V,E\backslash\{e\})}(v_i)||}{||R_{G(V,E)}(v_i)||}\,.$$

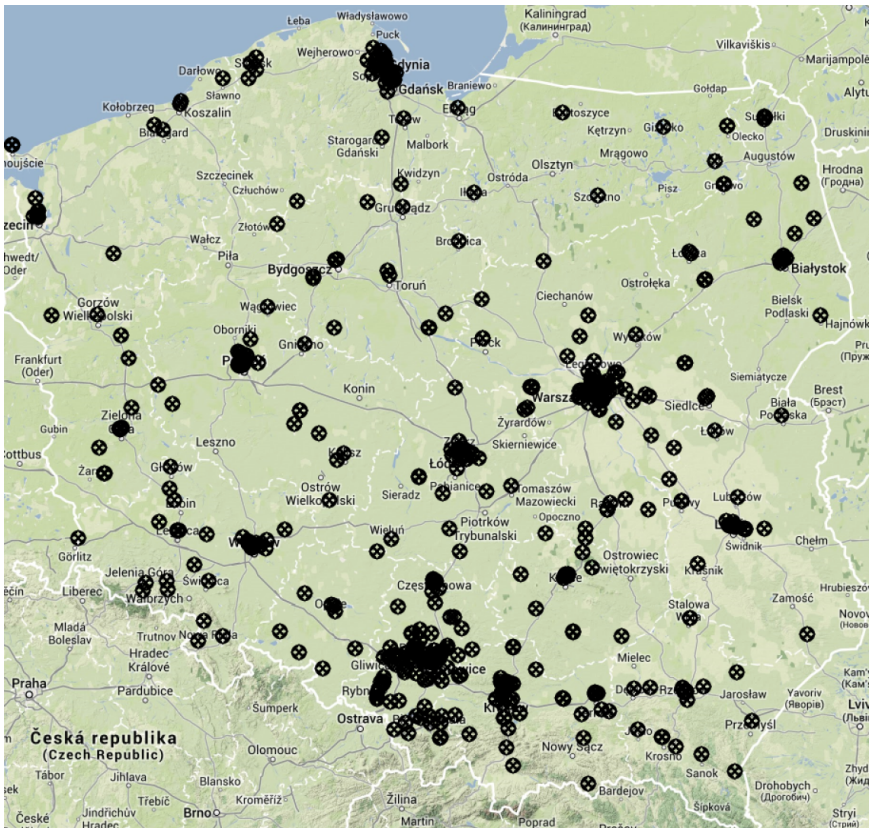
Analogicznie definiuję wskaźnik podatności u_{V} względem awarii pojedynczych AS. Oba wskaźniki umożliwiają ocenę odporności AS na awarie; przedstawione wyniki dowodzą, że większość polskich AS ma łącze zapasowe, ale można wskazać również dostawców bardzo słabo zabezpieczonych przed awarią.

Wprowadzam ponadto indeks istotności $s(\cdot)$ dla łącza będący liczbą par AS, które tracą łączącą je trasę BGP w wyniku awarii tego łącza:

$$s(e) = ||\{(v_i, v_{j\neq i}) : R_{G(V,E\setminus\{e\})}(v_i) \cap R_{G(V,E\setminus\{e\})}(v_j) = \emptyset\}|| \ .$$

Analogicznie definiuję istotność wierzchołków, czyli systemów autonomicznych.

Wprowadzenie specyficznych definicji i dodatkowe zabiegi wykonywane na grafie skierowanym mają na celu odtworzenie natury umów międzyoperatorskich, uniemożliwiających trasowanie ruchu przez AS leżące niżej w hierarchii dostawców (*tiered ISP model*). Przedstawiane osiągnięcie stanowi rozwinięcie istniejących, teoretycznych opracowań nt. odporności AS [13]. Moje badania pokazały, że w analizowanej rzeczywistej sieci nie ma łączy zupełnie redundantnych, oraz że rozkład istotności łączy jest bliski dwupunktowemu.



Rysunek 8. Lokalizacja AS według danych rejestrowych UKE (źródło: [H4], Fig. 3).

Pozyskanie dalszych danych, tj. lokalizacji geograficznych siedzib AS pozwoliło utożsamić je z lokalizacją urządzeń dostawców (szczególnie tych mniejszych). To z kolei umożliwia ocenę skutków hipotetycznego fizycznego ataku o założonym promieniu rażenia. Wyniki badań wskazują, że wartością graniczną promienia jest ok. 200 m — zniszczenie infrastruktury w większym promieniu powoduje lawinowy przyrost rozłączanych par systemów autonomicznych (por. [H4], Fig. 4).

Przeprowadzenie badań wymagało pozyskania tablic BGP. Adekwatnie do złożoności zaproponowanych algorytmów, ograniczono pobierane dane do systemów autonomicznych zarejestrowanych w Polsce. W celu ustalenia kraju rejestracji oraz dokładnego adresu siedziby podmiotu zarządzającego systemem autonomicznym, pobrano z UKE rejestr koncesji na działalność tego typu, por. rys. 8. Uzgodnienie nazw podmiotów zarządzających AS w obu rejestrach wymagało zaimplementowania autorskiej procedury dopasowania nazw.

Ocena związków pomiędzy istotnością kolokacji słownych a jej cechami sieciowymi

Wątek badania związków pomiędzy strukturą sieciową a zjawiskami zewnętrznymi, poruszony już w [H2]B, pojawił się również w ramach mojego udziału w projekcie HMO,[12] którego zasadniczy cel polegał na opracowaniu jak najlepszego modelu predykcji popytu na usługi hotelarskie. Zaproponowałem wzbogacenie modelu podstawowego bazującego na szeregach czasowych o moduł wyznaczający korekcję tego popytu w uzależnieniu od lokalnych wydarzeń publicznych. Kalendarz przeszłych i zaplanowanych wydarzeń udostępniany jest w większych miastach przez wydzieloną komórkę municypalną.[13] W zaproponowanym modelu korekcja popytu zależy od występowania w tytułach wydarzeń w danym dniu określonych kolokacji słownych. Oznacza to, że potencjalnie każda kolokacja języka wpływa w sposób liniowy na zwiększenie tego popytu (w praktyce ograniczam liczbę rozpatrywanych kolokacji). Istotnie, przypisanie kolokacjom wag oddziaływania wynikających z analizy wydarzeń historycznych spowodowało dalszą poprawę jakości predykcji. Moją ambicją badawczą stało się natomiast zbadanie związku pomiędzy wartościami otrzymanych wag a cechami kolokacji reprezentowanych jako połączenia w sieci słów języka.

Osiągnięciem opisanym w [H5] jest sprawdzenie hipotezy o **związku pomiędzy wagą kolokacji w modelu predykcyjnym a cechami topologicznymi kolokacji w sieci** słów w języku polskim. W takiej sieci węzły stanowią słowa a łącza — kolokacje słów. Sieć ma cechy sieci złożonej, co wynika z ekonomii językowej i zostało uznane w postaci prawa Zipfa [15]. Wytypowałem wektor czternastu cech topologicznych dla kolokacji (tj. par węzłów), które potencjalnie mogłyby mieć związek z jej wagą (wielkością skalarną) i rozwiązałem zadanie regresji liniowej, aby znaleźć najistotniejsze z nich. Eksperyment wykazał silne ujemne współczynniki regresji dla pośrednictwa kolokacji (*edge betweenness*) oraz dla maksimum ze stopni węzłów, por. tabela 2. W praktyce oznacza to, że kolokacje leżące na peryferiach sieci słów — te o mniej licznych powiązaniach, a więc bardziej spe- B

Tabela 2. Cechy topologiczne kolokacji $\{a, b\}$ słów a oraz b i odpowiadające im współczynniki regresji liniowej wag kolokacji w modelu predykcyjnym. Użyte oznaczenia wg. [30]: deg — stopień wierzchołka, C_{eig} — centralność wierzchołka wg analizy widmowej, C_{cen} — centralność wierzchołka ze względu na bliskość, C_{btw} — pośrednictwo wierzchołka (źródło: [H5], Table 6).

	Cecha	Wartość
1	Pośrednictwo krawędzi [7]	**-918,2**
2	Waga krawędzi, znormalizowana (liczba wystąpień kolokacji)	-5,4
3	$\deg(a) + \deg(b)$	-266,6
4	$\max(\deg(a), \deg(b))$	**-1408,0**
5	$\min(\deg(a), \deg(b))$	1141,4
6	$C_{\text{eig}}(a) + C_{\text{eig}}(b)$	-40,0
7	$\max(C_{\text{eig}}(a), C_{\text{eig}}(b))$	79,6
8	$\min(C_{\text{eig}}(a), C_{\text{eig}}(b))$	-119,5
9	$C_{\text{cen}}(a) + C_{\text{cen}}(b)$	0,9
10	$\max(C_{\text{cen}}(a), C_{\text{cen}}(b))$	-179,8
11	$\min(C_{\text{cen}}(a), C_{\text{cen}}(b))$	-178,9
12	$C_{\text{btw}}(a) + C_{\text{btw}}(b)$	-202,7
13	$\max(C_{\text{btw}}(a), C_{\text{btw}}(b))$	4,8
14	$\min(C_{\text{btw}}(a), C_{\text{btw}}(b))$	-197,9
15	wyraz wolny	8,3

[12] Projekt nr POIR.01.01.01-00-0050/15 finansowany przez EFRR, nazwa:*Hotels' Management Optimizer (HMO) — Pricing, Forecasting, Distribution — innowacyjne oprogramowanie nowej generacji do ustalania i prognozowania cen oraz zarządzania przychodami hoteli.* — w Wykazie osiągnięć jako pozycja [9.2].

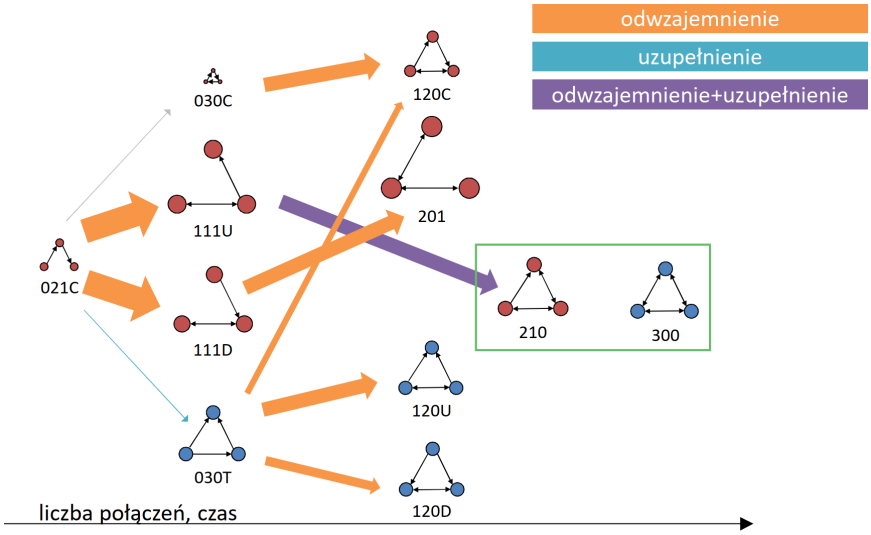
[13] Komórki te noszą różne polskie nazwy; jednolity odpowiednik angielski to *Convention Bureau*.

cyficzne — statystycznie częściej pojawiają się w tytułach wydarzeń o znacznym wpływie na popyt. Ta informacja może wspomóc w przyszłości typowanie obiecujących kolokacji do wykorzystania w modelu, zwłaszcza w sytuacji występowania szumu w postaci natłoku wydarzeń nieistotnych.

> Wymagającym warunkiem wstępnym przeprowadzenia tego badania było posiadanie zarówno danych o zajętości hoteli, jak i danych o wydarzeniach publicznych. Pozyskanie tych ostatnich w odpowiedniej objętości i jakości, dla różnych miast, było możliwe jedynie w drodze implementacji dedykowanego dla kolejnych miast oprogramowania monitorującego wydarzenia i pobierającego informacje ich dotyczące.

Model dynamiki ewolucji triad wykorzystujący równania kinetyczne reakcji chemicznej

Publikacja [H6] wprowadza nas w sferę badań zjawisk dynamicznych w sieci, tj. obejmujących zarówno zmiany struktury sieci, jak i zmiany właściwości sieci o ustalonej strukturze. Jednocześnie, podobnie jak w [H2] podejmuję tutaj próby weryfikacji istniejących i uznanych modeli fenomenów społecznych, wskazując przypadki, w których przestają one obowiązywać. Pierwszy z tych modeli mówi o stabilności strukturalnej triad przechodnich [18] i zakłada, że obecność takich triad świadczy o równowadze informacyjnej w grupie społecznej, a zatem jest pożądana i jest gwarantem stabilności takiej grupy. I odwrotnie, duży udział triad nieprzechodnich wiąże się z dyskomfortem jednostek, może prowadzić do agresji, samoagresji i, ogólnie, rychłej zmiany struktury sieci [5]. Weryfikując to twierdzenie posłużyłem się w badaniach migawkami fragmentu sieci użytkowników serwisu Instagram pozyskanymi w ramach pracy [19]. Przedmiotem mojego szczególnego zainteresowania była dynamika domykania się triady nieprzechodniej typu 210 do typu przechodniego 300 — czyli odwzajemnienia jedynej nieodwzajemnionej relacji obserwowania profili w grupie trzech użytkowników (por. rys. 9, triady w zielonej ramce).



Rysunek 9. Proces domykania się triady typu 021C w sieci użytkowników Instagrama. Wielkości triad reprezentują ich udział w ogóle obserwowanych triad. Grubości strzałek odpowiadają prawdopodobieństwu konkretnych scenariuszy domknięć. Ostatecznego domknięcia w triadę 300 nie zaznaczono strzałkami (źródło: [H1], Rys. 9.6).

C W [H6] proponuję **modelowanie dynamiki procesu domykania** triady typu 210 do triady 300 **za pomocą równania kinetycznego** reakcji chemicznej. Analiza czasów trwania triady 210 do momentu domknięcia wykazała, że proces ten jest analogią do reakcji chemicznej drugiego rzędu opisanej równaniem $\frac{1}{A} = \frac{1}{A_0} + kt$, gdzie A oznacza tempo reakcji domykania wszystkich takich triad w sieci, malejące w czasie t ze współczynnikiem k. Uznanie takiego modelu parametrycznego

tranzycji triad otwiera nowe możliwości analizy dynamiki, m.in. wyszukiwania anomalii w tym procesie. Anomalie takie zostały odnalezione i przedyskutowane w polskiej, rozszerzonej wersji artykułu;[14] stawiam tam hipotezę o ich związku z terminarzem ferii zimowych.

Przeprowadzenie tego badania wymagało, oprócz specyficznego algorytmu pobierania danych, opisanego w rozdz. 4.3.5, szczególnego zaprojektowania relacyjnej bazy danych. Schemat danych został zoptymalizowany pod względem wyszukiwania triad określonego typu, zakodowanego w postaci zagnieżdżonych zapytań SQL.

Model epidemiczny dynamiki popularności wiadomości w serwisie typu «social news»

Osiągnięte w [H6] wyniki dobitnie uzmysławiają, że operator serwisu społecznościowego może w istotny sposób ingerować w naturalną dynamikę procesów społecznych, np. stymulując użytkowników do odwzajemnienia relacji bycia obserwowanym lub polubionym (w znaczeniu definiowanym przez serwis). Tendencja ingerowania operatorów w zachowania społeczne nasiliła się do tego stopnia, że obecnie rozważane jest ograniczanie jej w drodze działań legislacyjnych. Istnieją więc uzasadnione przesłanki kontynuacji tego wątku badawczego. Moje dalsze prace skupiły się na analizie dynamiki popularności wiadomości zamieszczanych w serwisie typu *social news*, tj. takim, którego społeczność użytkowników troszczy się zarówno o wyszukiwanie interesujących wiadomości, jak i o ich ocenę, która przekłada się na pozycję wiadomości w obrębie witryny. W toku badań [H7] weryfikuję kolejną popularną opinię [11] o adekwatności modelu epidemicznego SIR (*susceptible–infectious–recovered*) w modelowaniu dyfuzji wiadomości (a więc i jej popularności). Model ten zakłada, że tempo "zarażania" (susceptible→infectious) jest proporcjonalne do liczby aktualnie zarażonych — co niewątpliwie ma miejsce dla zakażeń drogą kontaktu fizycznego. Ta koncepcja jest powszechnie przenoszona na grunt kontaktów wirtualnych i tam rozwijana o elementy specyficzne, ale przy zachowaniu powyższej liniowej zależności. To ostatnie wynika z zakładania, że wiadomość dyfunduje poprzez powiązania w sieci użytkowników, bez szerszej publicznej ekspozycji.

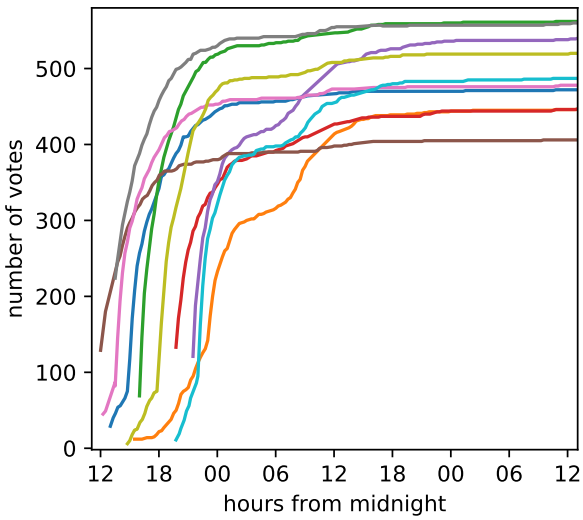
W [H7] wykazuję, że **dyfuzja informacji wśród użytkowników «social news» przebiega inaczej**. Wykresy łącznego zainteresowania wiadomościami przypominają odpowiedź członu inercyjnego na skok jednostkowy. Niekiedy przebieg ten bywa poprzedzony fazą liniową. Przykładowe wykresy przedstawia rys. 10, przy czym przyjętą miarą popularności jest liczba głosów oddanych na wiadomość przez społeczność serwisu. Proponuję następujący model dynamiki liczby oddanych głosów v na wiadomość pojawiającą się w chwili t_0:

$$\dot{v} = \begin{cases} \alpha & \text{dla } t - t_0 < \tau \\ \beta r(t)(N - v)^{\xi} & \text{w p.p. ,} \end{cases}$$

gdzie $[\alpha, \tau, \beta, N, \xi]$ to parametry modelu: nachylenie fazy liniowej, czas do przejścia w fazę nieliniową, wzmocnienie i wartość asymptoty oraz współczynnik kształtu fazy nieliniowej. Funkcja $r(\cdot)$ jest uśrednionym profilem aktywności dobowej ogółu użytkowników serwisu.

Model taki dokładnie dopasowuje się do dynamiki poszczególnych wiadomości, a jego struktura i wyniki dogłębniejszych analiz świadczą o tym, że sieć indywidualnych powiązań użytkownika zamieszczającego wiadomość nie ma istotnego wpływu na dynamikę ani w fazie liniowej, ani w nieliniowej. Wszystkie wiadomości opublikowane na stronie głównej serwisu doświadczają lawinowego wzrostu popularności, w sposób oczywisty niezależnego od struktury sieci użytkowników serwisu. Niektóre z nich przechodzą również początkową liniową fazę wzrostu, nie znajdującą uzasadnienia w strukturze kontaktów znalazcy wiadomości. Inne wykryte korelacje pomiędzy wartościami parametrów modelu i sieci okazały się słabe. Również próby badania związku parametrów modelu

[14] M. Kamola, *Dynamika triad w serwisie Instagram*, Przegląd Telekomunikacyjny s. 1174-1178, DOI: 10.15199/59.2016.8-9.71, 8–9/2016 — w Wykazie osiągnięć jako pozycja [4.9].



Rysunek 10. Popularność wybranych wiadomości w serwisie wykop.pl w funkcji czasu. Obserwujemy: początkową opcjonalną fazę liniową, sezonowość dobową (nocny spadek dynamiki), nasycenie na zróżnicowanych poziomach, dynamikę wzrostu popularności nieskorelowaną z poziomem nasycenia (źródło: [H7], Fig. 1b).

z występowaniem słów kluczowych w treści wiadomości nie doprowadziły do wyników umożliwiających dobrą predykcję dynamiki poszczególnych doniesień.

Badania wykonano dla danych z polskiego serwisu wykop.pl. Pozyskanie niezbędnych danych wymagało zaimplementowania programu systematycznie i nieprzerwanie analizującego strony www serwisu. Prace analityczne przeprowadzono za pomocą autorskich programów prototypowych, korzystających z bibliotek numerycznych, optymalizacyjnych, analizy sieci i komputerowej analizy języka naturalnego.

4.3.5 Odkrywanie i przekształcanie sieci

W tym rozdziale opisuję prace wykonane w obrębie trzech wątków badawczych dotyczących konstruowania nowych sieci i przekształcania sieci istniejących, przy wykorzystaniu wiedzy dziedzinowej, założeniu o naturze sieci wynikowej oraz uwzględnieniu prywatności informacji — odpowiednio.

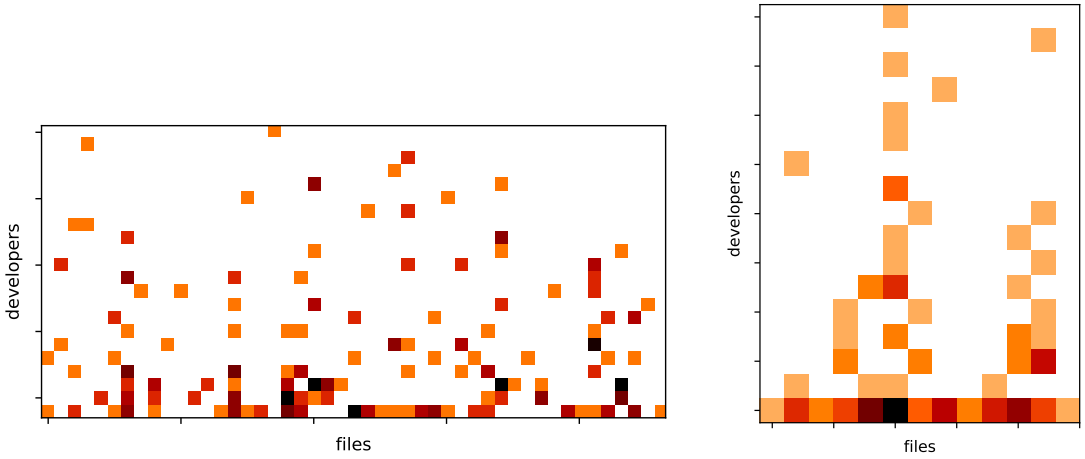
Metodyka tworzenia sieci współpracy programistów i sieci zależności modułów oprogramowania

Rekonstruowanie struktur sieciowych z danych surowych wymaga z reguły wykorzystania wiedzy z dziedziny, której te dane dotyczą. Tę specyfikę demonstruje publikacja [H2] — aby dokonać analizy własności sieci, przedstawionej w rozdz. 4.3.4, należy te sieci odtworzyć. Użyte wyjściowe ślady cyfrowe opisują aktywność programistów i obejmują kod źródłowy projektów. Aby przełożyć je na uniwersalny język opisu sieci, musimy wykorzystać wiedzę praktyczną o procesie tworzenia oprogramowania: faktoryzacji i optymalizacji kodu źródłowego, rozgałęziania i łączenia wersji, rejestrowania zmian w kodzie wykonanych przez konkretnych programistów itd.

D W [H2] proponuję **metodę rekonstrukcji sieci modułów oprogramowania i powiązań zawodowych programistów** z danych dostępnych w repozytorium projektów o otwartym kodzie źródłowym. Odtworzenie sieci zależności modułów stanowi rozwinięcie statycznej analizy kodu przedstawionej w [H3] i zaprezentowanej na rys. 6. Uwzględniłem praktykę grupowania powiązanych logicznie modułów w tym samym katalogu i wprowadziłem wagi w sieci powiązań odpowiada-

jące odległości modułów w drzewie katalogów kodu źródłowego. Metoda ta daje wyniki podobne do stosowanej dotychczas semantycznej analizy kodu [20], ale nie wymaga eksperckiego wglądu w szczegóły projektu w celu dostrojenia algorytmu, tj. np. utworzenia specyficznej listy słów pospolitych (*stop words*). Natomiast sieć powiązań między programistami powstała poprzez ważoną projekcję w grafie dwudzielnym programista-aktywność.

Uzyskane sieci obu typów wykazują cechy sieci złożonych. Rys. 11 przedstawia przykładowe macierze sąsiedztwa sieci dwudzielnych aktywności programistów w plikach źródłowych (tzw. *code commits*). Uzyskane wyniki ujawniają bezskalowy charakter sieci programistów i kodu. Oznacza



a) b)

Rysunek 11. Intensywność działań poszczególnych programistów na plikach źródłowych w projektach a) TensorFlow i b) AirBnB (źródło: [H2], Fig. 1b i 1f).

to, że prawie w każdym projekcie pojawiają się "superprogramiści", modyfikujący większość plików źródłowych, jaki i "supermoduły", nad którymi pracuje większość programistów. Zjawiska te utrzymują się nawet po skrupulatnej weryfikacji danych i wyeliminowaniu artefaktów (pliki dokumentacji, noty prawne, działalność botów). Bezskalowy charakter obu sieci stoi w ewidentnej sprzeczności z prawem Conwaya, postulującym enkapsulację funkcjonalności kodu i działalności programistycznej w mniej więcej zrównoważonych grupach. Mimo to analizowane projekty są w pełni udane, co może oznaczać, że 1) prawo Conwaya nie obowiązuje dla zwirtualizowanych organizacji wytwórczych lub 2) dane wejściowe są niepełne lub błędne, lub 3) w przypadku pracy zdalnej należy przyjąć inne definicje grupy oraz inne definicje modułu.

> Badania wykonano dla danych z dziewięciu repozytoriów kodu źródłowego i aktywności programistów. Łącznie poddano analizie za pomocą oprogramowania Madge i własnego około połowę z ponad 70 tys. pobranych plików z kodem źródłowym.

Wyniki badań [H2], a także [H6] oraz [H7] rozpatrywane łącznie uzmysławiają, że znane od lat zjawiska dotyczące sieci społecznych, silnie uargumentowane socjologicznie i mające odbicie w uznanych modelach zdają się w pewnych warunkach nie obowiązywać. Warunki te to istnienie nowych reguł lub kanałów komunikacji: wykorzystywania narzędzi wirtualizujących pracę zespołową programistów [H2], zachęt do odwzajemniania relacji [H6] czy wypromowania wiadomości na szczyt listy [H7].

Dwa kolejne przykłady odkrywania sieci wykorzystują wiedzę z dziedziny komputerowej analizy języka naturalnego. Zagadnienie to jest pozornie odległe od tematyki osiągnięcia naukowego — niemniej język naturalny generuje sieci złożone zarówno w wyniku analizy syntaktycznej (współ-

występowania słów), jak i semantycznej (związków pomiędzy pojęciami). Przykład ich wykorzystania przedstawiłem już w [H5]. Komputerowa analiza języka stała się dominującym narzędziem ekstrakcji informacji z zasobów Internetu mimo rosnącej liczby interfejsów programistycznych udostępniających dane ustrukturyzowane i mimo wcześniejszych wysiłków spopularyzowania etykietowania stron www znacznikami semantycznymi ([H1], s. 75 i nast.). Komputerowa analiza języka przeżywa obecnie rozkwit, wykorzystując szeroko techniki uczenia maszynowego, których zastosowanie stało się możliwe dzięki dostępności bardzo wielu dokumentów w postaci cyfrowej oraz popularności komunikatorów tekstowych. Nie należy zapominać, że obok pomyślnych adaptacji technik uczenia maszynowego do obróbki tekstów, wiodące ośrodki naukowe — w tym polskie — wykonały i udostępniają narzędzia i zasoby słownikowe opracowane przez lingwistów w oparciu o wielopokoleniową wiedzę ekspercką.

Metodyka wyboru istotnych kolokacji słów w dokumencie tekstowym

Typowym zadaniem NLP jest odnajdowanie dokumentów podobnych pod względem zawartości. Sieć podobieństwa dokumentów możemy zrekonstruować poprzez projekcję w sieci dwudzielnej, w której dokumenty są traktowane jak węzły typu A, a słowa w dokumentach lub kolokacje słów — jak węzły typu B. Wykorzystanie kolokacji, czyli sekwencji dwusłowowych wynika z kompromisu pomiędzy poszukiwaniem atrybutów dokumentu, które byłyby już precyzyjne, ale jeszcze nie egzotyczne. Pojedyncze słowa są często wieloznaczne, natomiast sekwencje trzech słów i dłuższe we współczesnych korpusach pojawiają się zbyt rzadko, by opierać na nich NLP. Niezależnie od wyboru długości sekwencji, częstość występowania atrybutów, np. kolokacji w korpusie tekstów ma rozkład Zipfa [15]. Rozkład ten jest bliskim odpowiednikiem rozkładu potęgowego; w szczególności jest on bezskalowy, co utrudnia wskazanie zbioru kolokacji tworzącego bazę dla dalszych porównań dokumentów, np. według podobieństwa kosinusowego. Praktykuje się ograniczanie zbioru kolokacji do tych najistotniejszych, przy czym badania wskazują [26], że spośród licznych technik ekstrakcji terminów istotnych bardzo dobre wyniki daje indeks Jaccarda, wskazujący w jakim stopniu pokrywają się zbiory dokumentów zawierających oba słowa składowe kolokacji. Dobór kolokacji determinuje strukturę sieci dokumentów otrzymanej w wyniku projekcji.

D W [H1], s. 148 proponuję **metodykę eksperckiego wyboru istotnych kolokacji** dla korpusu dokumentów tekstowych. Podobnie jak w osiągnięciach [H1](A) oraz [H3], kolokacje są prezentowane na płaszczyźnie dwóch kryteriów: a) *powszechności* występowania w korpusie oraz b) *specyficzności* obliczanej jako indeks Jaccarda dla słów składowych. Ponownie, ostateczny wybór kolokacji znaczących pozostawiany jest ocenie eksperckiej. Natomiast wstępny zbiór potencjalnie znaczących kolokacji C jest wyznaczany przez autorski algorytm operujący na dwóch rankingach kolokacji: (a_i) oraz (b_i), sporządzonych dla obu kryteriów, odpowiednio. Algorytm rozpoczyna działanie od przeglądu obu rankingów, aż do *znalezienia pierwszej kolokacji* występującej wysoko w obu rankingach. Kolejne kolokacje *dołączane są* do C poprzez cykliczny przegląd obu rankingów w rozszerzanym zakresie, począwszy od miejsca znalezienia pierwszej kolokacji — aż do uzyskania zbioru o założonej liczności N. Algorytm przedstawiono schematycznie poniżej:

- Dane wejściowe: (a_i), (b_i), N
- Dane wyjściowe: C
- Znalezienie pierwszej kolokacji: $n := 0$; Powtarzaj: $n := n+1, C := \{a_1, \ldots a_n\} \bigcap \{b_1, \ldots b_n\}$ dopóki $C = \emptyset$
- Dołączanie kolejnych kolokacji: $k := 0$; Powtarzaj:
  - $k := k + 1, \quad a^\star := (a_{n-k}, \ldots, a_{n+k}), \quad b^\star := (b_{n-k}, \ldots, b_{n+k})$
  - $C := C \bigcup (\{a^\star\} \bigcap \{b^\star\})$; jeśli $||C|| \geq N$, zakończ, zwracając C.

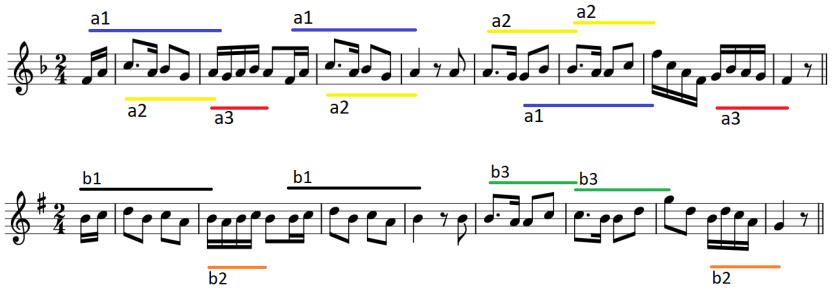
Metoda tworzenia sieci utworów muzycznych

Istniejące i oczywiste analogie w terminologii pojęć języka naturalnego i języka muzyki stały się motywacją do badań przedstawionych w publikacji [H8]. Prace w dziedzinie komputerowej analizy muzyki trwają nieprzerwanie; są one istotne zarówno dla muzykologów, jak i z punktu widzenia komercyjnego, np. w celu trafnego doboru list odtwarzania, zgodnych z indywidualnymi gustami muzycznymi. We współpracy z mgr Barbarą Laskowską [23] poszukiwaliśmy muzycznych odpowiedników słów, zwrotów i całych zdań tekstu. Badania literatury wykazały, że pojęciem tyle powszechnym, ile niedookreślonym w analizie muzycznej jest *motyw muzyczny*. Specjaliści różnią się, już od początku XX w. [34], w ocenie, które cechy utworu są najbardziej charakterystyczne i składają się na pojęcie motywu.

W dalszych pracach spośród cech uznawanych współcześnie [22] za charakterystyczne dla melodii jednogłosowych, wybraliśmy i zaadaptowaliśmy cztery: odległości diatoniczne i chromatyczne kolejnych nut utworu, kontur melodii oraz informację o długości nut, którą wyraziliśmy jako stosunek długości sąsiadujących nut. Dokładniejsza analiza problemu doprowadziła do wniosku, że analogie pomiędzy językiem naturalnym a językiem muzyki są powierzchowne i dotyczą jedynie terminologii, gdyż konstrukcja przekazu muzycznego podlega zupełnie innym regułom niż składnia języka. Wiąże się ona z wielowymiarową, zmysłową percepcją treści muzycznej. Wnioski te skłoniły nas do poszukiwania numerycznej reprezentacji utworów adekwatnej do specyfiki odbioru muzyki, a jednocześnie w pełni wytłumaczalnej.

W [H8] proponujemy zdefiniowanie motywu w melodii na bazie cech jej fragmentów, analogicznie do n-znakowych fragmentów tekstu nazywanych n-gramami. W komputerowej analizie języka fragmenty te służą do zbudowania n-gramowego modelu statystycznego języka. Analizując utwór, również wyodrębniamy wszystkie n-gramy cech charakterystycznych melodii, a następnie konstruujemy motywy jako zbiory pewnych n-gramów. Elementy zbioru nazywamy *realizacjami motywu*; są nimi n-gramy dostatecznie podobne do siebie według zaproponowanej miary podobieństwa uwzględniającej praktykę muzykologiczną. Obliczając podobieństwo motywów, używamy indeksu Jaccarda dla realizacji motywów. Natomiast podobieństwo dwóch utworów jest równe maksimum podobieństwa par motywów pochodzących z każdego z utworów.

Wartość n generująca n-gramy jest parametrem algorytmu; w połączeniu z charakterem zaproponowanych miar generuje ona siatkę podobieństwa utworów, która bardziej przypomina topologię sieci niż przestrzeń euklidesową. Osiągnięciem badawczym w [H8] jest więc **opracowanie metody tworzącej sieć utworów muzycznych**. Praktyczną użyteczność metody sprawdzono znajdując utwory podobne za pomocą standardowych dla sieci metod grupowania aglomeracyjnego, uzyskując zadowalającą zgodność grupowania z oryginalnym sklasyfikowaniem utworów w korpusie. D



Rysunek 12. Fragmenty dwóch melodii ludowych; realizacje pojedynczego motywu w utworze oznaczono tym samym kolorem. Motywy a2 i b3 są podobne, ponieważ zawierają jedną identyczną realizację — na początku 5. pełnego taktu (źródło: [H8], Fig. 3).

Metoda znajduje również nieoczywiste podobieństwa pomiędzy utworami z różnych klas, dające się wytłumaczyć na gruncie analizy muzykologicznej. Przykłady motywów wyodrębnionych w dwóch utworach przedstawiono na rys. 12.

Kolejny ważny wątek dotyczy procesu rekonstruowania sieci ukierunkowanego na otrzymanie struktury o określonych własnościach. Opracowując taki algorytm rekonstrukcji, podobnie jak w przypadkach omówionych powyżej, dostarczamy nowej i użytecznej wiedzy o sieci — lecz tym razem nie wynika ona z naszej specyficznej, technicznej wiedzy o naturze procesów tworzących sieć, lecz z ostatecznego, konkretnego celu analitycznego, dla którego sieć ta jest tworzona. Wyszczególnienie takiego celu pozwala na bardziej liberalne podejście do rekonstrukcji sieci: nie musi ona modelować dokładnie relacji wszystkich, a jedynie szczególne. Liberalizacja podejścia jest niezbędna, by sprostać pewnym ograniczeniom zewnętrznym. W rzeczywistości, ograniczenia te są warunkami wstępnymi konstruowania metod: możemy zrekonstruować sieć tylko w takiej formie i pozwolić sobie następnie na tylko takie analizy, jakie wynikają z ograniczeń pierwotnych.

W mojej praktyce napotkałem dwa takie przypadki. Pierwszy dotyczy rekonstrukcji sieci przy ograniczeniu ilościowym dotyczącym zakresu jej eksplorowania. Ograniczenia ilościowe wynikają albo z naturalnego braku danych pomiarowych, albo z polityki ograniczania ich dostępności przez źródło, z różnych przyczyn wewnętrznych. Próba pokonania tych ograniczeń może skutkować znacznym kosztem finansowym (dokupienie danych lub budowa infrastruktury technicznej do ich zbierania), dodatkowym czasem, zwiększonym ryzykiem (zdrowotnym, biznesowym) itp. Jeśli pozostajemy przy ograniczonym eksplorowaniu, to zrekonstruowana sieć zawsze pozostanie niekompletna. Istnieją cele analityczne (np. oszacowanie rozkładu stopni węzłów, rozkładu współczynnika gronowania), dla których już niewielka próbka danych daje dobre przybliżenie wskaźników całej sieci [24]. Istnieją jednak aspekty sieci szczególnie wrażliwe na ubytek danych, np. jej średnica oraz spójność.

Metoda rekonstrukcji sieci przy ograniczeniach zasobowych

Strategia próbkowania sieci nie jest obojętna dla jakości uzyskanych wyników. W publikacji [H6] przedstawiam algorytm opracowany wspólnie z mgr. Jakubem Jarzyńskim w ramach pracy magisterskiej pod moją opieką [19], realizujący zadanie rekonstrukcji grafu powiązań użytkowników serwisu Instagram przy ograniczeniu na liczbę danych o profilach użytkowników udostępnianych w jednostce czasu. Ponieważ ostatecznie graf ten posłużył do modelowania dynamiki triad, por. [H6]C, celem opracowanego algorytmu jest wydobycie podgrafu o maksymalnej gęstości.

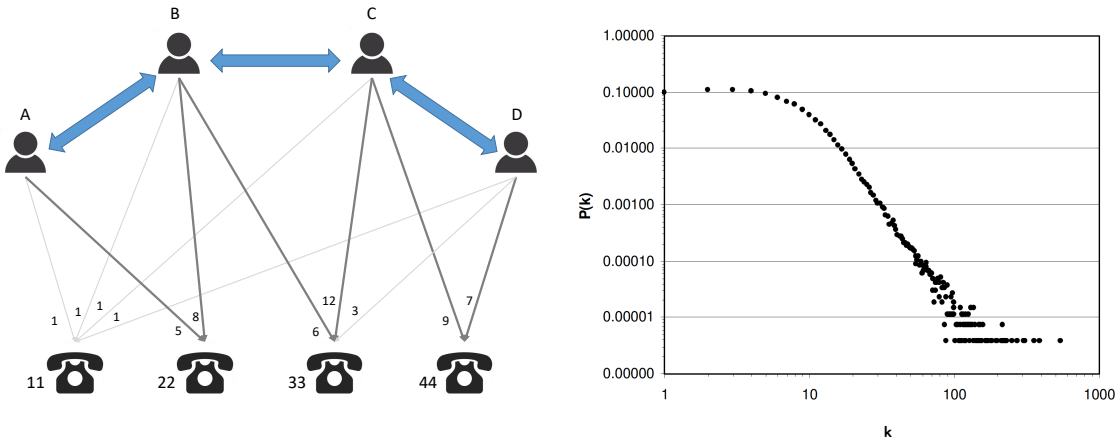
E Przedstawiona w [H6] metoda **rekonstrukcji gęstego grafu przy ograniczeniach zasobowych** polega na iteracyjnym typowaniu do eksplorowania takich węzłów-profili użytkowników, które byłyby w maksymalnym stopniu połączone z węzłami już pozyskanymi od dostawcy danych. Jest ona modyfikacją poszukiwania wszerz z jednoczesnym sortowaniem nowoznalezionych węzłów ze względu na liczbę ich połączeń z dotychczasowo pozyskanymi węzłami. W każdej iteracji budowany podgraf jest powiększany tylko o N najsilniej usieciowionych nowoodkrytych węzłów. Pomijane są przy tym węzły o wysokim, większym od M, stopniu. W przypadku węzłów-profili użytkowników portalu społecznościowego i przy właściwym doborze parametrów M i N, metoda wykazała oczekiwaną tendencję do odnajdowania użytkowników z tego samego miasta.

Przedstawiony algorytm powstał w wyniku wielu eksperymentów pobierania danych z serwisu przy niejawnych ograniczeniach na liczbę zapytań. Taktyka nieujawniania przez dostawcę limitów liczby zapytań w komunikacji algorytmu z serwisem jest powszechna i staje się dodatkową przeszkodą w interakcji z dostawcami danych. Opracowania algorytmu w pełni wykorzystującego limity zapytań dokonuje się w drodze prób i błędów — praktyka ta jest tyleż nieelegancka co skuteczna i powszechna, a przedstawiony algorytm jest tego przykładem.

Procedura projekcji ukierunkowana na otrzymanie sieci złożonej

Drugi przypadek rekonstrukcji obejmuje ograniczenia jakościowe danych. W pracy [H9] rozwiązuję problem odtworzenia sieci powiązań społecznych użytkowników indywidualnych telefonii stacjonarnej na podstawie rejestru rozmów (*call detail records*, CDR). Wyjściowy zbiór CDR jest ułomny, bowiem identyfikatory klientów wykonujących rozmowy oraz wybierane przez nich numery zostały zaszyfrowane. Poszczególne połączenia telefoniczne można więc utożsamiać co najwyżej z połączeniami węzłów w sieci dwudzielnej, z numerami telefonów jako węzłami górnymi, a abonentami jako węzłami dolnymi. Zadanie polega na opracowaniu metody rekonstrukcji powiązań między abonentami.

Praktycznej użyteczności tego zadania nie należy ograniczać wyłącznie do informatyki śledczej. Można bowiem wyobrazić sobie liczne scenariusze, w których operator telekomunikacyjny udostępnia je rozmyślnie w formie okrojonej w trosce o ochronę prywatności abonentów, jednocześnie zlecając podmiotowi zewnętrznemu analizę pewnych interesujących go cech sieci — jak np. wielkości mikrospołeczności albo współczynników gronowania. Są one niezbędne w planowaniu strategicznym, np. w konstrukcji taryf i promocji. Operator również wykorzystuje takie informacje strukturalne do prognozowania liczby rezygnacji z usług [16]. Wydelegowanie takich zadań (i zanonimizowanych danych) do podmiotu zewnętrznego może również wynikać z ich rozmiaru, który wymaga wykorzystania specjalistycznego sprzętu do analizy.[15] Niezależnie od konkretnych zastosowań w telekomunikacji, rekonstrukcja właściwej sieci połączeń z danych niepełnych stanowi obecnie powszechny problem analityczny: wnioskujemy o relacjach biznesowych, politycznych, ścieżkach ruchu itp. na podstawie wspólnych atrybutów przedsiębiorstw, polityków, pojazdów, odpowiednio. Z koniecznością odtwarzania docelowej struktury połączeń spotykałem się w większości projektów realizowanych przez studentów przedmiotu *Techniki analizy sieci społecznych*, przedstawionego w rozdz. 6.1.



a) b)

Rysunek 13. a) Zastosowana projekcja w sieci dwudzielnej danych CDR dla $r = 5$. Szarymi strzałkami oznaczono połączenia abonentów z numerami; w projekcji pomijamy połączenia wykonane mniej niż r razy w miesiącu. Grube błękitne strzałki oznaczają zrekonstruowane powiązania między abonentami. b) Rozkład stopni wierzchołków-abonentów w zrekonstruowanej sieci (źródło: [H9], Fig. 2).

[15] Opisywane prace wymagały wykonania masywnych obliczeń równoległych na maszynie z 96 rdzeniami i 60 GB RAM (realia z r. 2010).

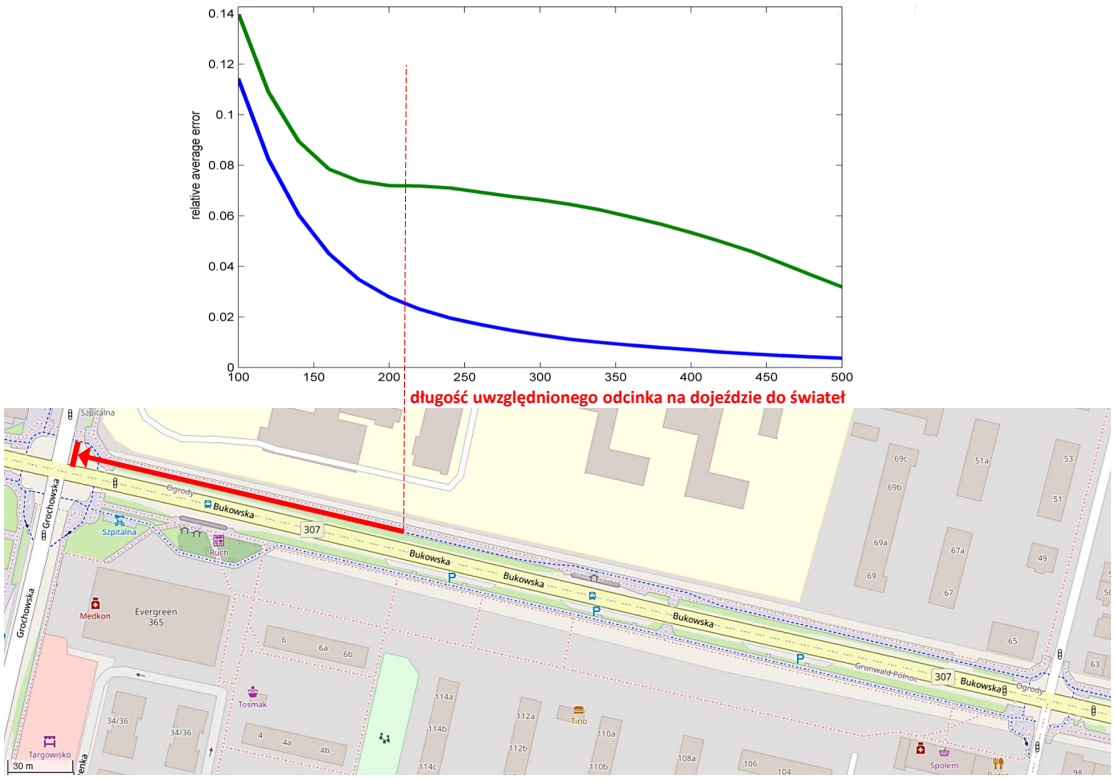
E F W [H9] proponuję sposób **projekcji w sieci dwudzielnej prowadzący do tego, że zrekonstruowana sieć posiada większość cech sieci złożonej** charakteryzującej kontakty społeczne, ponieważ jej węzły reprezentują indywidualnych abonentów telekomunikacyjnych. Statystyki liczby wykonanych połączeń uprawniają do przyjęcia takiego założenia. Skądinąd wiadomo [29], że projekcja prosta w sieciach dwudzielnych prowadzi przy określonych założeniach do powstania sieci bezskalowej. Zaproponowany przeze mnie algorytm projekcji jest wariantem projekcji prostej, przy czym postuluję ograniczenie liczby uwzględnianych węzłów dolnych do r najczęściej wybieranych numerów docelowych przez każdego abonenta. Takie ograniczenie jest zgodne z charakterem większości relacji społecznych, ale również znacząco przyspiesza wykonanie projekcji dla dużych sieci. Zrekonstruowana sieć zachowuje bezskalowość dla dość szerokiego zakresu r. Co więcej, analiza dynamiki nowych klientów w tak zrekonstruowanej sieci jest w pełni zgodna z uznanym modelem preferencyjnego dołączania dla sieci złożonych [4]. Zasadę działania i wyniki algorytmu przedstawia rys. 13.

Procedura konstrukcji użytecznego grafu anonimowego

Powyższy przykład wprowadził nas w wątek uwzględniania prywatności danych jako istotnego ograniczenia lub wręcz celu tworzenia bądź transformacji sieci. W [H9] wyszliśmy w zadaniu rekonstrukcji grafu od spreparowanych, zanonimizowanych danych dostarczonych przez operatora. W [H10] rozwiązuję zadanie odwrotne: posiadając dokładne ślady cyfrowe położenia pojazdów (czyli występując w roli ich dostawcy), należy spreparować na ich podstawie taką sieć, która będzie użyteczna w określonym zakresie dla podmiotu zewnętrznego, jednocześnie chroniąc prywatność kierowców i wykonywanych przez nich podróży. Waga problemu skutecznej ochrony prywatności danych lokalizacyjnych jest od dawna znana [3]; wysoka unikatowość śladów cyfrowych powoduje, że zwykła anonimizacja identyfikatorów nie wystarcza. Stosowane są więc różne metody zaszumiania, mieszania, agregacji a nawet wtórnej generacji danych lokalizacyjnych w celu uniemożliwienia ustalenia tożsamości ich właścicieli [33].

F W [H10] proponuję **procedurę naniesienia tras przejazdów na sieć ulic** w taki sposób, aby chroniąc anonimowość kierowców, pozostawić jednocześnie dane istotne do oceny ich stylu jazdy np. przez służby albo ubezpieczycieli. Wynikowa sieć ulic pozbawiana jest informacji geoprzestrzennych, jak również informacji o odległości pomiędzy węzłami-skrzyżowaniami. Taki szczególny izomorfizm sieci oryginalnej i wynikowej powoduje, że zadanie deanonimizacji węzłów-skrzyżowań staje się NP-zupełne [14]. Zaproponowane podejście jest pozornie sprzeczne z postulatem udostępnienia informacji o stylu jazdy kierowców, który przecież wynika z ich położenia i prędkości. Jednak analiza danych rzeczywistych wykazuje, że styl ten wyraża się głównie na dojeździe do skrzyżowań, por. rys. 14. Oznacza to, że można bez dużej straty pominąć początkowe fragmenty odcinków. ujednolicając tym samym długości krawędzi w grafie wynikowym do założonej wartości, np. 200 m i kodując lokalizację pojazdu względem kolejnego skrzyżowania. Zaproponowana metoda nie wyklucza wdrożenia dodatkowych zabiegów ochrony prywatności kierowców w przypadkach szczególnie łatwych do zdeanonimizowania topologii ulic.

Przedstawione prace wymagały implementacji wydajnego algorytmu projekcji śladów GPS na siatkę ulic tak, aby skorygować naturalnie występujące zakłócenia pomiarowe. Pomocne w tym okazało się wykorzystanie bazy danych z indeksami geoprzestrzennymi, co umożliwiło uwzględnienie w projekcji tylko ulic z najbliższego otoczenia konkretnego śladu.



Rysunek 14. Błędy szacowania maksymalnej prędkości odcinkowej (ziel.) oraz maksymalnego przyspieszania lub hamowania (nieb.) w funkcji długości końcowego fragmentu odcinka branego pod uwagę w szacowaniu (na podst. [H10], Fig. 2 i 3).

4.3.6 Podsumowanie

W czasach łatwej dostępności danych (niekiedy bardzo obfitych) i popularyzacji algorytmów ich analizy (niekiedy bardzo zaawansowanych) istotne jest, aby nie utracić z oczu sensu modelowanego zjawiska, oraz aby trafnie wybierać spośród gotowych metod.

Wola rozumienia zjawiska najczęściej skutkuje powstaniem nowych wartościowych modeli interdyscyplinarnych. Ich działanie jest zrozumiałe, w przeciwieństwie do złożonych modeli bazujących na uczeniu maszynowym, w przypadku których dopiero od niedawna kładzie się nacisk na możliwość objaśnienia wyników. Moje istotne z tej perspektywy osiągnięcia wykorzystują wiedzę z zakresu automatyki [H7], fizyki [H6], muzykologii [H8], inżynierii ruchu [H10] i telekomunikacji [H9], [H4] w konstruowaniu i analizie sieci złożonych. Część prac w sposób naturalny została zrealizowana we współpracy ze znawcami tych dziedzin — w części wykorzystuję moje własne doświadczenie.

Wybór gotowych metod analizy oraz poruszanie się w wielowymiarowej przestrzeni parametrów ich pracy wymagają systematyzacji podejścia. Moim osiągnięciem jest zaproponowanie metodyk oceny i porównania wyników działania gotowych algorytmów analizy sieci złożonych. Dotyczą one zarówno prac studialnych [H1], [H2], jak i pracy z gotowym zastosowaniem praktycznym [H3], realizowanej w ramach dużego projektu badawczo-rozwojowego.

Wszystkie badania składające się na przedstawione osiągnięcie zostały przeprowadzone dla danych rzeczywistych, z ukierunkowaniem na potencjalne i faktyczne zastosowania praktyczne. Wynalezione, zaadaptowane lub wykorzystane przeze mnie metody i podejścia były każdorazowo uwarunkowane wprost ilością i jakością dostępnych danych. Ponieważ w wielu przypadkach badane hipotezy dotyczyły zjawisk społecznych ([H2], [H6]–[H10]), zostały one zweryfikowane w drodze eksperymentów obliczeniowych przeprowadzonych dla odpowiednio dużych zbiorów danych.

5 Informacje o pozostałej istotnej aktywności naukowej

Tematyka i charakter zrealizowanych przeze mnie pozostałych prac badawczych i rozwojowych mają ścisły związek z charakterem instytucji, w których jestem zatrudniony lub z którymi współpracowałem, i w niektórych przypadkach stanowią kontynuację osiągnięć z pracy doktorskiej [P13]. Zostały one zaprezentowane poniżej w podziale tematycznym.

5.1 Inżynieria ruchu w sieci pakietowej

Wykonywane badania dotyczyły trzech aspektów inżynierii ruchu w sieciach pakietowych: jakości, ekonomii i energooszczędności usług przesyłu danych. W projekcie *Platforma budowy usług multimedialnych*[16] badałem możliwości techniczne strumieniowania wideo w wysokiej rozdzielczości pomiędzy stacjami VectaStar, oferującymi bezprzewodową transmisję w technologii ATM. Wyniki pomiarów i ich omówienie zostały przedstawione w publikacji [P1].

Znajomość praktyczna technik priorytetyzacji ruchu w sieci ułatwiła realizację wątku badawczego dotyczącego ekonomicznych aspektów dynamicznego kontraktowania usług przez klientów o zróżnicowanych funkcjach użyteczności pasma dla aplikacji elastycznych [P2]. Zaproponowałem architekturę systemu adaptacyjnego przydziału pasma w odpowiedzi na sygnały zwrotne od użytkowników. Wykonane prace dobrze wpisywały się w nurt ówczesnych trendów badawczych, dążących do dynamicznej wyceny usług i szukania alternatyw dla ściśle hierarchicznej struktury operatorów internetowych. Wraz ze współpracownikiem odniosłem się do tego ostatniego postulatu, opracowując propozycję skalowalnej architektury kontraktacji usług, przedyskutowaną w [P3]. Dopuszcza ona istnienie wielu punktów kontraktowania ruchu, co umożliwia zarówno dekompozycję zadania rezerwacji pasma w skali globalnej, adekwatnie do obecnej topologii, jak i współzawodniczenie usługodawców kontraktowania. Publikacja zawiera studium opłacalności takiego przedsięwzięcia. Ponadto, w ramach projektu *Usługi i sieci teleinformatyczne następnej generacji*[17] współpracowałem przy budowie prototypu systemu badawczego — uniwersalnego symulatora mechanizmów aukcyjnych, umożliwiającego algorytmom-agentom użytkownika komunikowanie się w ustalonym schemacie w celu osiągnięcia równowagi rynkowej, a następnie przydziału zasobów sieciowych. Rozwiązanie zostało omówione w [P4].

Wątek badawczy poświęcony energooszczędnemu sterowaniu siecią realizowany był w ramach projektu Econet.[18] Praca zespołowa skupiała się głównie na zagadnieniu opracowania adekwatnego modelu sieci z protokołem IP, uwzględniającego aktualne i przyszłe możliwości sprzętowe ograniczania zużycia energii przy niepełnym zapotrzebowaniu na przepustowość. Wytworzony model

[16] Projekt nr WKP_1/1.4.1/1/2006/125/125/682/2007 finansowany przez EFRR, nazwa: *Platforma budowy usług multimedialnych nowej generacji dla sieci komputerowych i mobilnych* — w Wykazie osiągnięć jako pozycja [9.7].

[17] Projekt nr PBZ-MNiSW-02/II/2007 finansowany przez MNiSW, nazwa: *Usługi i sieci teleinformatyczne następnej generacji – aspekty techniczne, aplikacyjne i rynkowe* — w Wykazie osiągnięć jako pozycja [9.6].

[18] Projekt nr INFSO-ICT-258454 finansowany w 7PR KE, nazwa: *Econet (Low Energy Consumption Networks)* — w Wykazie osiągnięć jako pozycja [9.5].

prowadził do powstania mieszanego zadania programowania w celu znalezienia optymalnej energetycznie konfiguracji sieci, przy uwzględnieniu aktualnego zapotrzebowania oraz ograniczeń na jakość świadczonych usług. Został on przedstawiony w publikacjach [P5]–[P8]. Z realizacją projektu związane jest również moje autorskie osiągnięcie [P9], prezentowane w języku polskim w [P10]. Przedstawiam tam propozycje heurystycznych algorytmów oszczędzających energię w dwóch przykładowych sieciach szkieletowych. Algorytmy te mają uzupełniać lukę pomiędzy istniejącymi podejściami skrajnymi:

- sterowaniem opartym na ocenie wyłącznie lokalnego obciążenia łączy w routerze [6],
- sterowaniem scentralizowanym wykorzystującym informacje globalne o zapotrzebowaniu na przepustowość [8].

Zauważmy przy tym, że na zużycie energii na łączu składa się w uproszczeniu a) część stała, niezależna od natężenia ruchu, oraz b) część zależna (np. proporcjonalnie) od natężenia ruchu. O ile strategie energooszczędnego zarządzania siecią są w przypadkach skrajnych (tylko a albo tylko b) proste do skonstruowania na drodze analitycznej, o tyle w przypadkach pośrednich i najczęstszych prowadzą do zadań programowania mieszanego o dużej złożoności [31]. Proponuję i badam trzy algorytmy aktywacji, w sytuacji nadmiernego wzrostu ruchu na łączu e_l, aktualnie wyłączonych łączy: 1) włączenie ostatnio wyłączonego, 2) włączenie tego, które najbardziej zmniejszy liczbę najkrótszych ścieżek bieżąco prowadzących przez e_l, 3) włączenie tego, które najbardziej zmniejszy ruchu przez e_l, przy znanej macierzy zapotrzebowania na przepustowość (*traffic matrix*, TM). Komplementarnie, proponuję cztery strategie dezaktywacji łączy w sytuacji spadku ruchu. Można wyłączyć łącze: A) najmniej obciążone, B) przez które przechodzi najmniej najkrótszych ścieżek, C) które doprowadzi do najmniejszego możliwego wzrostu sumarycznej długości ścieżek, D) które doprowadzi do najmniejszego możliwego wzrostu sumarycznego ruchu — przy znajomości macierzy zapotrzebowania. Wyniki badań wskazują jednoznacznie, że stosunkowo złożone strategie: 2 w połączeniu z B lub C, wykorzystujące informacje o trasowaniu w sensie ilościowym (trasy ścieżek), a nie jakościowym (TM), działają najgorzej. Oznacza to, że stosowanie zaawansowanych algorytmów decyzyjnych w sieci wymaga dostarczenia adekwatnie kompletnych i dokładnych danych.

Oprócz powyższych badań, moje prace badawcze w projekcie Econet obejmowały stworzenie laboratorium pomiarowego zużycia energii przez urządzenia sieciowe, z autorskim systemem zbierania i przetwarzania danych pomiarowych. Laboratorium odegrało fundamentalną rolę w osiągnięciu zasadniczych celów naukowych projektu, służąc również kontynuacji badań [21]. Architektura i rola laboratorium zostały przedstawione w publikacjach [P11] oraz [P12].

Publikacje

[P1] M. Kamola, P. Arabas, *Wykorzystanie technologii Vecta Star do przekazu audiowizualnego wysokiej rozdzielczości*, Przegląd Telekomunikacyjny, s. 1508–1513, 8–9/2009[19]
- **4 P** MNiSW

[P2] P. Paulski, M. Kamola, *Optimal Bandwidth Allocation in IP Network; the case of QoS-sensitive user utility functions*, Intl. Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), IEEE Explore, ISBN: 978-1-56555-320-0, 2008[20]
- **3 P** MNiSW

[19] W Wykazie osiągnięć jako pozycja [4.23]
[20] W Wykazie osiągnięć jako pozycja [7.9]

[P3] M. Kamola, P. Arabas, *Dynamically Established Transmission Paths in the Future Internet - Proposal of a Framework*, Biuletyn Techniczny PAN (59.3) s. 357–366, DOI: 10.2478/v10175-011-0043-9, 2011[21]
- **30 P** MNiSW

[P4] M. Kamola, M. Karpowicz, K. Malinowski, E. Niewiadomska-Szynkiewicz, *System badawczy*, w: Mechanizmy aukcyjne i giełdowe w handlu zasobami telekomunikacyjnymi, red. J. Lubacz, WKiŁ, s. 126–138, ISBN: 978-83-206-1825-9, 2011[22]
- **4 P** MNiSW

[P5] R. Bolla, R. Bruschi, F. Davoli, P. Lago, Á. Bakay, R. Grosso, M. Kamola, M. Karpowicz, L. Koch, D. Levi, G. Parladori, D. Suino, *Large-scale validation and benchmarking of a network of power-conservative systems using ETSI's Green Abstraction Layer*, Transactions on Emerging Telecommunications Technologies (27.3) s. 451–468, DOI: 10.1002/ett.3006, 2016[23]
- **25 P** MNiSW

[P6] E. Niewiadomska-Szynkiewicz, A. Sikora, P. Arabas, M. Kamola, M. Mincer, J. Kołodziej, *Dynamic power management in energy-aware computer networks and data intensive computing systems*, Future Generation Computer Systems (37) s. 284–296, DOI: 10.1016/j.future.2013.10.002, 2014[24]
- **40 P** MNiSW, JCR **IF**=2,786

[P7] M. Kamola, E. Niewiadomska-Szynkiewicz, P. Arabas, A. Sikora *Energy-saving Algorithms for the Control of Backbone Networks: A Survey*, Journal of Telecommunications and Information Technology, s. 13–20, 2/2016[25]
- **12 P** MNiSW

[P8] E. Niewiadomska-Szynkiewicz, A. Sikora, P. Arabas, M. Kamola, K. Malinowski, P. Jaskóła, M. Marks, *Network-Wide Power Management in Computer Networks*, ITC Specialist Seminar on Energy Efficient and Green Networking (SSEEGN), IEEE Explore, DOI: 10.1109/SSEEGN.2013.6705398, 2013[26]
- **10 P** MNiSW, nagroda **best paper award**

[P9] M. Kamola, P. Arabas, *Shortest Path Green Routing and the Importance of Traffic Matrix Knowledge*, Tyrrhenian Intl. Workshop on Digital Communications — Green ICT (TIWDC), IEEE Explore, DOI: 10.1109/TIWDC.2013.6664215, 2013[27]
- **15 P** MNiSW, WoS=5, Scopus=11, GS=14

[P10] M. Kamola, *Energooszczędne trasowanie ruchu w sieci IP z OSPF*, Przegląd Telekomunikacyjny s. 1049–1055, 1–3/2014[28]
- **9 P** MNiSW

[P11] M. Kamola, P. Arabas, P. Jaskóła, E. Niewiadomska-Szynkiewicz, K. Malinowski, M. Karpowicz, A. Sikora, M. Mincer, M. Marks, *ECONET – energooszczędne sieci IP*, Przegląd Telekomunikacyjny, s. 964–970, 8–9/2013[29]

[21] W Wykazie osiągnięć jako pozycja [4.18]
[22] W Wykazie osiągnięć jako pozycja [2.5]
[23] W Wykazie osiągnięć jako pozycja [4.7]
[24] W Wykazie osiągnięć jako pozycja [4.12]
[25] W Wykazie osiągnięć jako pozycja [4.8]
[26] W Wykazie osiągnięć jako pozycja [7.7]
[27] W Wykazie osiągnięć jako pozycja [7.6]
[28] W Wykazie osiągnięć jako pozycja [4.13]
[29] W Wykazie osiągnięć jako pozycja [4.14]

[P12] E. Niewiadomska-Szynkiewicz, P. Arabas, M. Kamola, K. Malinowski, T. Wiśniewski, *Sterowanie energooszczędną siecią teleinformatyczną*, w: Aktualne problemy automatyki i robotyki, red. K. Malinowski, J. Józefczyk, J. Świątek, Exit, s. 505–514, ISBN: 978-83-7837-040-6, 2014[30]
- **5 P** MNiSW

5.2 Analiza danych i sterowanie obiektami przemysłowymi

Ten obszar badawczo-rozwojowy stanowi kontynuację badań wykonanych w ramach pracy doktorskiej, której istotę stanowił problem sterowania systemem modelowanym przez symulator numeryczny. Z uwagi na nieznaną, niewypukłą dziedzinę modelu, we wcześniejszych badaniach opracowałem procedurę projektowania systemu sterowania, a w szczególności doboru i adaptacji bezgradientowych algorytmów optymalizacji sprzęganych z symulatorem. W odróżnieniu od modeli rozpatrywanych w mojej rozprawie [P13], podjęte później prace dotyczyły fragmentu sieci gazowej wysokiego ciśnienia firmy Gaz-System SA zarządzanej przez oddział we Wrocławiu. Jako osoba odpowiedzialna za adaptację istniejącego systemu automatycznego sterowania tłoczniami gazu, posłużyłem się autorską procedurą projektową, wprowadzając liczne modernizacje istniejącego sprzężenia symulatora z optymalizatorem. Wyniki przedstawiłem w [P14]. Napotkanym zupełnie nowym zjawiskiem, nieuwzględnionym w rozprawie, był dualizm funkcjonowania symulatora SIMONE, zdolnego albo obliczać stan ustalony sieci, albo prowadzić symulację dynamiczną. Rozwiązanie tego i wielu innych inżynierskich wyzwań związanych m.in. z bezpieczeństwem i niezawodnością systemu uważam za istotne osiągnięcie zawodowe i cenne dopełnienie aktywności stricte badawczej.

W tym samym obszarze lokuje się moja współpraca z firmą Neptis SA, operatorem aplikacji Yanosik, zorientowana na analizę śladów GPS pojazdów. Część zadań skutkująca analizą sieci drogowej została przedstawiona w [H10]. Inne opublikowane prace [P15] dotyczą modelowania macierzy ruchu pomiędzy obszarami aglomeracji. Do badania zależności natężenia ruchu i czynników pogodowych stosuję analizę składowych głównych (*principal component analysis*, PCA).

Prace badawcze w zakresie przetwarzania danych masowych w czasie rzeczywistym mają swoją kontynuację w ramach rozpoczętego we wrześniu 2020 r. i trwającego projektu telemetrii oglądalności kanałów telewizyjnych, realizowanego na zlecenie Krajowej Rady Radiofonii i Telewizji.[31] Moje zadania polegają m.in. na wyszukiwaniu zależności w strumieniu danych, grupowaniu danych, projektowaniu algorytmów analitycznych w wersji docelowej oraz definiowaniu dla nich interfejsów programistycznych.

Publikacje

[P13] M. Kamola, *Algorithms for Optimisation Problems with Implicit and Feasibility Constraints*, rozprawa doktorska na Wydziale Elektroniki i Technik Informacyjnych PW, 2004[32]

[P14] M. Kamola, S. Plamowski, Cz. Godlejewski, K. Antoniewicz, A. Gromnicki, *Problem optymalnego tłoczenia gazu w sieci wysokiego ciśnienia*, Nafta-Gaz, s. 24–32, 1/2015[33]
- **6 P** MNiSW

[P15] M. Kamola, Jakub Wesołowski, *Techniki analizy i modelowanie więźby ruchu miejskiego*, Transport Miejski i Regionalny, s. 12–18, 2/2019[34]
- **5 P** MNiSW

[30] W Wykazie osiągnięć jako pozycja [2.3]

[31] Jest to projekt badawczo-rozwojowy finansowany ze środków KRRiT — w Wykazie osiągnięć jako pozycja [15.1].

[32] W Wykazie osiągnięć jako pozycja [1.3]

[33] W Wykazie osiągnięć jako pozycja [4.10]

[34] W Wykazie osiągnięć jako pozycja [4.5]

5.3 Badania i rozwój potencjału rynku domenowego

Unikatowa pozycja NASK PIB jako operatora krajowego rejestru DNS wiąże się z dużym poczuciem odpowiedzialności i troski tej instytucji o właściwe wykorzystanie możliwości jakie daje i jakie jeszcze dać może ta technologia. Moje prace badawcze i rozwojowe odzwierciedlają to podejście. W [P16] prezentuję analizę zapytań do rejestru WHOIS (informacje o zarządcach domen). Wyniki wskazują na istotny związek popularności serwisów z liczbą zapytań WHOIS o odpowiadające im nazwy domen — dokładne lub przybliżone.

Odkryta w [P16] natura komercyjnego poszukiwania wolnych nazw domen uzmysławia odpowiedzialność krajowych rejestrów za kształt rynku domenowego. Moim udziałem w tym procesie było wytworzenie autorskiego oprogramowania wyszukiwarki wolnych nazw w domenie *.pl*. Algorytm realizuje poszukiwanie semantyczne, uwzględniając zarówno odległość znaczeniową synonimów wprowadzonych słów kluczowych, jak i ich częstość występowania w korpusie języka polskiego. Prototyp oprogramowania został udostępniony krajowym rejestratorom DNS w partnerskim programie oceny. Zapoczątkował on również moje dalsze prace nad przetwarzaniem języka naturalnego [P17], [P18].

Kolejną zaproponowaną innowacją jest wykorzystanie DNS jako powszechnego rejestru urządzeń inteligentnych w tzw. Internecie Rzeczy (*Internet of Things*, IoT). Rozszerzenia specyfikacji DNS pod nazwą DNSSec oraz DANE[35] umożliwiają zastąpienie tradycyjnego łańcucha poświadczeń tożsamości rozwiązaniem alternatywnym. W [P19] uzupełniam istniejące propozycje [35] o wykorzystanie alternatywnych, tanich i powszechnych dostawców tożsamości. W zaimplementowanym rozwiązaniu integruję typowy serwer DNSSec oraz Profil Zaufany (pz.gov.pl) w roli "lekkiego" dostawcy tożsamości, w pełni funkcjonalny prototyp systemu. Wykonane prace stanowiły inspirację i wkład merytoryczny do wniosków o projekty badawcze w ramach programu Horyzont 2020,[36] których NASK i Politechnika Warszawska były uczestnikami.

Publikacje

[P16] M. Kamola, *Who is asking and for what: WHOIS traffic analysis*, Journal of Telecommunications and Information Technology, s. 14–21, 4/2012[37]
- **7 P** MNiSW

[P17] K. Ciecierski, M. Kamola, *Comparison of Text Classification Methods for Government Documents*, International Conference on Artificial Intelligence and Soft Computing, LNCS 12415, s. 39–49, DOI: 10.1007/978-3-030-61401-0_4, 2020[38]
- **20 p** MNiSW

[P18] M. Kamola, *Analytics of Industrial Operational Data Inspired by Natural Language Processing*, IEEE Intl. Congress on Big Data, IEEE Explore, DOI: 10.1109/BigDataCongress.2015.108, 2015[39]
- **10 P** MNiSW

[35] *DNS Security Extensions* oraz *DNS-based Authentication of Named Entities*, standardy określone w RFC 4033 oraz 6698.

[36] Wniosek o projekt DINET w konkursie H2020-SU-ICT-2018-2020 oraz wniosek o projekt DYNAMIC w konkursie H2020-SU-DS-2018-2019-2020.

[37] W Wykazie osiągnięć jako pozycja [4.15]

[38] W Wykazie osiągnięć jako pozycja [2.2]

[39] W Wykazie osiągnięć jako pozycja [7.4]

[P19] M. Kamola, *Internet of Things with Lightweight Identities Implemented Using DNS DANE - Architecture Proposal*, Sensors (8), DOI: 10.3390/s18082517, 2018[40]
- **100 p** MNiSW, JCR **IF**=3,031

5.4 Prace rozwojowe dla cyberbezpieczeństwa

Zadania realizowane przeze mnie w obszarze cyberbezpieczeństwa początkowo miały charakter implementacyjny. W projekcie *Opracowanie systemu informatycznego umożliwiającego digitalizację...*[41] implementuję archiwizację dokumentów w tzw. drzewach Merklego przechowywanych w chmurze Amazon.

Kolejne zadanie ma charakter koncepcyjny i polega na zaprojektowaniu funkcjonalności i architektury systemu ankietowania podmiotów świadczących usługi kluczowe i cyfrowe, w celu odtworzenia sieci powiązań pomiędzy usługami. Zadanie było realizowane w ramach projektu NPC.[42] Za istotne osiągnięcie uważam zaproponowaną strukturę i schemat obiegu informacji pomiędzy centrum ankietowania a podmiotami, umożliwiające utożsamienie różnych nazw tych samych usług (używanych przez różnych ankietowanych), lecz bez niepotrzebnej ingerencji w tajemnicę handlową. W ramach tego samego projektu realizowałem również badania stricte naukowe skutkujące publikacją [H3].

5.5 Analiza danych heterogenicznych

Problematyka pozyskiwania, łączenia, wnioskowania i prezentacji danych z różnych źródeł i w różnych formatach jest uniwersalna i dostosowuje się ją do konkretnego zadania analitycznego. W projekcie stypendialnym pn. CONTENT 1.0 stanąłem przed problemem zaprojektowania schematu pozyskiwania i analizy wzmianek (odniesień w tekście do słów kluczowych — ang. *mentions*) w sposób możliwie uniezależniony od konkretnego źródła, ale jednocześnie pozwalający na ich łączną ocenę. Za osiągnięcie uważam zaproponowanie bazowej struktury dokumentów tekstowych, specyfikację języka zapytań stanowiącego znaczne rozwinięcie wyrażeń regularnych, a także opracowanie metodyki tabelarycznej wizualizacji wyników, pozwalającej na wyznaczanie wielu statystyk, jak również wygodny eksport danych. Osiągnięcia te zostały przedstawione w [P20].

Publikacje

[P20] M. Tanaś, M. Kamola, R. Lange, M. Fila, *BigData w edukacji. CONTENT 1.0 — prototyp aplikacji do analizy treści internetu*, Wydawnictwo APS, ISBN: 978-83-66010-29-1, 2019[43]
- **20 p** MNiSW

5.6 Narzędzia i algorytmy przetwarzania równoległego i rozproszonego

Moje prace w tej dziedzinie początkowo miały charakter studialny i popularyzatorski [P21], prowadząc do opracowania własnych narzędzi i algorytmów. W [P22] przedstawiam i oceniam własne środowisko programowania rozproszonego i heterogenicznego dla pakietów Matlab i Octave, wykorzystujące koncepcję pamięci wspólnej. Możliwości wykorzystania wykonanego w kontekście

[40] W Wykazie osiągnięć jako pozycja [4.6]

[41] Projekt nr DOBR/0071/R/ID1/2012/03 finansowany przez NCBiR, nazwa: *Opracowanie systemu informatycznego umożliwiającego digitalizację, wieczystą archiwizację, zarządzanie i bezpieczne udostępnianie w formie elektronicznej dokumentów i materiałów archiwalnych* — w Wykazie osiągnięć jako pozycja [9.3].

[42] Projekt nr CYBERSECIDENT/369195/I/NCBR/2017 finansowany przez NCBiR, nazwa: *Narodowa Platforma Cyberbezpieczeństwa* (NPC) — w Wykazie osiągnięć jako pozycja [9.1]

[43] W Wykazie osiągnięć jako pozycja [1.1]

projektu *Usługi i sieci teleinformatyczne następnej generacji...*[44] symulatora rozproszonego do obliczeń numerycznych przedstawiam natomiast w [P23]. Symulator, pierwotnie zaprojektowany jako symulator giełdy zasobów sieciowych, z powodzeniem może służyć do rozproszonego rozwiązywania zadań kombinatorycznych i innych.

Publikacje

[P21] M. Kamola, J. Błaszczyk, B. Kubica, E. Niewiadomska-Szynkiewicz, *Programowanie rozproszone w środowiskach sieciowych oparte na wywołaniach zdalnych procedur*, w: Programowanie równoległe i rozproszone, red. A. Karbowski, E. Niewiadomska-Szynkiewicz, Oficyna Wydawnicza Politechniki Warszawskiej, s. 234–346, ISBN 978-83-7207-803-2, 2009[45]
- **3 P** MNiSW

[P22] M. Kamola, *Shared Memory for Matlab and Octave: Another Package for Distributed Programming*, IFAC Symposium on Large Scale Systems Theory and Applications (40.9) s. 316–321, DOI:
10.3182/20070723-3-PL-2917.00051, 2007[46]
- **3 P** MNiSW

[P23] M. Kamola, *Software Environment for Market Balancing Mechanisms Development, and Its Application to Solving More General Problems in Parallel Way*, International Workshop on Applied Parallel Computing, LNCS 7133, s. 231–241, ISBN: 978-3-642-28150-1, DOI: 10.1007/978-3-642-28151-8_23, 2012[47]

6 Osiągnięcia dydaktyczne, organizacyjne oraz popularyzatorskie

Poniżej przedstawiam działania tworzące środowisko dla moich osiągnięć naukowych.

6.1 Dydaktyka i opieka naukowa

Uważam, że współpraca ze studentami realizowana poprzez zajęcia dydaktyczne oraz indywidualną opiekę naukową stanowi cenny element otoczenia, w którym naukowiec realizuje własne badania. Inspiruje, motywuje, ćwiczy umiejętności organizacyjne, a często również weryfikuje idee.

Moja aktywność dydaktyczna na Wydziale Elektroniki i Informatyki Stosowanej Politechniki Warszawskiej obejmuje wszystkie etapy kształcenia. W ramach zatrudnienia jestem angażowany w realizację następujących przedmiotów:

- na studiach I stopnia:
  - Programowanie zdarzeniowe — wykład
  - Systemy operacyjne — laboratorium
  - Programowanie obiektowe — laboratorium
  - Programowanie w języku C — laboratorium

[44] Projekt nr PBZ-MNiSW-02/II/2007 finansowany przez MNiSW, nazwa: *Usługi i sieci teleinformatyczne następnej generacji – aspekty techniczne, aplikacyjne i rynkowe* — w Wykazie osiągnięć jako pozycja 9.7.

[45] W Wykazie osiągnięć jako pozycja [2.6]

[46] W Wykazie osiągnięć jako pozycja [7.10]

[47] W Wykazie osiągnięć jako pozycja [2.4]

- na studiach II stopnia:
  - Techniki analizy sieci społecznych — wykład, projekt
  - Sieci i sterowanie systemami — wykład
  - Sterowanie sieciami komputerowymi — wykład
  - Podstawy obliczeń równoległych i rozproszonych — projekt
  - Sieci komputerowe — laboratorium
  - Computer networks — laboratorium
- na studiach podyplomowych Zarządzanie Zasobami IT:
  - Klasyczne metodyki zarządzania projektami — wykład

Oto krótkie omówienie najistotniejszych ze względu na mój wkład, przedmiotów na poszczególnych stopniach. Od 2008 r. prowadzę wykłady i koordynuję przedmiot *Programowanie zdarzeniowe*, realizowany na studiach inżynierskich na kierunku Informatyka. W ramach wykładów, zasadniczo poświęconych wykorzystaniu języka Java w komunikacji zdarzeniowej, prezentuję wraz ze współpracownikiem aktualne problemy i trendy inżynierii oprogramowania związane np. z rozwojem wzorców projektowych i gotowych architektur (*frameworks*), integracją z API usług sieciowych czy bezpieczeństwem i podatnościami kodu źródłowego. W ramach przedmiotu realizowane były projekty współorganizowane z działem informatyki jednego z największych światowych banków inwestycyjnych.

Od 2015 r. wraz ze współpracownikiem prowadzę wykłady i zajęcia projektowe z przedmiotu *Techniki analizy sieci społecznych*, którego uruchomienia byłem inicjatorem. Kurs ten należy do grupy obieralnych przedmiotów technicznych na studiach 2. stopnia i nieprzerwanie cieszy się kompletem słuchaczy. Łączy wiedzę teoretyczną o sieciach złożonych z technikami pozyskiwania, analizy i prezentacji danych. Elementem wyróżniającym przedmiot są zajęcia projektowe, w ramach których staramy się oferować tematy związane z aktualnymi problemami społecznymi i aktualnie dostępnym danymi publicznymi (np. dotyczącymi mobilności ludzi, powiązań organizacji, sieci opinii). Wiele zagadnień projektowych zostało rozwiniętych do formy tematów dyplomowych, a zdobyte doświadczenia wpłynęły istotnie na fakt powstania i zawartość monografii [H1].

Od 2010 r. prowadzę wykłady z zarządzania projektami informatycznymi według metodyk klasycznych w ramach Studium Podyplomowego *Zarządzanie zasobami IT* przy WEiTI PW. Wykład obejmuje kompleksowo otoczenie projektowe, tj. metodyki PRINCE2 i PMI, jak również pokrewne im metodyki zarządzania programem i portfelem, a także, przekrojowo, rolę Biura Projektów w organizacji. Treść wykładu była wielokrotnie aktualizowana wg zmian w metodykach oraz własnych doświadczeń zawodowych i prac badawczych (m.in. w związku z badaniami przedstawionymi w [H2]).

Pod moją opieką powstało i zostało obronionych 18 prac magisterskich oraz 23 inżynierskie.

Wraz ze współpracownikiem przygotowałem nowy przedmiot obieralny II stopnia pt. *Sieci inteligentnych urządzeń*, zaakceptowany programowo i przeznaczony do uruchomienia w semestrze letnim 2021 r. Zakres przedmiotu obejmuje analizę systemów cyber-fizycznych oraz urządzeń IoT w połączeniu z rozproszonym zastosowaniem zaawansowanych algorytmów uczenia maszynowego. Przedmiot będzie zawierał część projektową.

6.2 Działalność organizacyjna i udział w dyskursie naukowym

Działania o charakterze zarządczym i organizacyjnym:

- Zastępca kierownika Pracowni Sterowania Siecią w NASK, 2010–2016.
- Pełnienie obowiązków członka-założyciela z ramienia NASK Polskiej Platformy Technologii Informatycznych, 2006–2007.
- Zasiadanie w Radzie Naukowej NASK jbr IV kadencji, 2010–2011.
- Kierownik zespołu NASK w projekcie ECONET (7. Program Ramowy UE, 16 konsorcjantów).
- Kierownik zespołu NASK w projekcie *Platforma budowy usług multimedialnych* (Program EFRR, 3 konsorcjantów).
- Kierownik techniczny zespołu Politechniki Warszawskiej w projekcie QOSIPS (5. Program Ramowy UE, 5 konsorcjantów).
- Udział w pracach zespołowych w opracowaniu strategii informatyzacji dla Urzędu Miasta St. Warszawy, 2006

Udział w dyskursie naukowym:

- Współpraca naukowa z Uniwersytetem w Genui w następstwie projektu Econet,[18] skutkująca m.in. wykonaniem recenzji rozprawy doktorskiej i udziałem w komisji egzaminacyjnej.
- Udział w komitecie organizacyjnym Tyrrhenian International Workshop on Digital Communications — Green ICT (2013 r.)
- Przygotowanie recenzji artykułów naukowych i referatów konferencyjnych m.in. dla *Int. J. Appl. Math. Comput.*, *Foundations of Computing and Decision Sciences*, *MDPI Sensors*, *MDPI Electronics*, *Journal of Telecommunications and Information Technology* — lista recenzji w rozdz. II.13 Wykazu osiągnięć.
- Udział w opracowaniu strategii rozwoju sztucznej inteligencji przez Ministerstwo Cyfryzacji w Grupie tematycznej 2: *Finansowanie badań i rozwoju* (2018 r.).
- Opracowanie opinii dotyczącej programu studiów 1. stopnia na kierunku *Cyberbezpieczeństwo* dla WEiTI PW (2019 r.)
- Przygotowanie recenzji kilkudziesięciu prac dyplomowych na WEiTI PW.

6.3 Działalność popularyzatorska i otrzymane nagrody

Działalność popularyzatorska:

- Prezentacja istoty głębokiego uczenia maszynowego w cyklu konferencji *Perspektywy dla rozwoju Internetu Rzeczy — Samorząd Przyszłości* organizowanych przez Ministerstwo Cyfryzacji i Kancelarię Prezydenta RP, 2020.
- Prezentacja osiągnięć Pracowni Sterowania Siecią na konferencji Secure, 2014.

- Prezentacje wyszukiwarki wolnych nazw domen na cyklicznych konferencjach biznesowych organizowanych przez NASK dla rejestratorów domen, 2013, 2015.
- Artykuł *Tomografia sieciowa* w Biuletynie NASK, 2007.

Otrzymane nagrody i wyróżnienia:

- Nagroda naukowa zespołowa I stopnia za współautorstwo książki *Programowanie Równoległe i Rozproszone* [P21], 2009 .
- Nagroda zespołowa III stopnia Rektora PW za osiągnięcia dydaktyczne w r. akad. 2014/2015, 2016.
- Nagroda zespołowa III stopnia Rektora PW za osiągnięcia w dziedzinie informatyki oraz automatyki i robotyki — za udział w projekcie 5. Programu Ramowego UE *Quality of Service and Pricing Differentiation for IP Services* (QOSIPS), 2003.
- Nagroda *best paper award* za publikację [P8], 2013.
- Nominacja do nagrody *best paper award* za publikację [H7], 2020.

Literatura

[1] Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, and Athanasios V. Vasilakos. The role of big data analytics in internet of things. *Computer Networks*, 129:459 – 471, 2017. Special Issue on 5G Wireless Networks for IoT and Body Sensors.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web*, pages 181–190, 2007.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] Peter S Bearman and James Moody. Suicide and friendships among american adolescents. *American journal of public health*, 94(1):89–95, 2004.

[6] Aruna Prem Bianzino, Luca Chiaraviglio, Marco Mellia, and Jean-Louis Rougier. GRiDA: GReen Distributed Algorithm for energy-efficient IP backbone networks. *Computer Networks*, 56(14):3219–3232, 2012.

[7] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[8] Antonio Cianfrani, Vincenzo Eramo, Marco Listanti, and Marco Polverini. An OSPF enhancement for energy saving in IP networks. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 325–330. IEEE, 2011.

[9] R. Cohen and S. Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.

[10] Melvin E Conway. How do committees invent. *Datamation*, 14(4):28–31, 1968.

[11] Daryl J Daley and David G Kendall. Epidemics and rumours. *Nature*, 204(4963):1118–1118, 1964.

[12] N. Davies and E. Tabakowska. *Europa: rozprawa historyka z historią*. ZNAK, 2004.

[13] Wenping Deng, Merkouris Karaliopoulos, Wolfgang Mühlbauer, Peidong Zhu, Xicheng Lu, and Bernhard Plattner. k-fault tolerance of the internet as graph. *Computer Networks*, 55(10):2492 – 2503, 2011.

[14] David Eppstein. Subgraph isomorphism in planar graphs and related problems. *CoRR*, cs.DS/9911003, 1999.

[15] Zipf George. Human behavior and the principle of least effort, 1949.

[16] Witold Gruszczyński and Piotr Arabas. Application of social network inferred data to churn modeling in telecoms. *Journal of Telecommunications and Information Technology*, (2):77–86, 2016.

[17] Frank Harary, Robert Zane Norman, and Dorwin Cartwright. *Structural models: An introduction to the theory of directed graphs*. Wiley, 1965.

[18] Paul W Holland and Samuel Leinhardt. Local structure in social networks. *Sociological methodology*, 7:1–45, 1976.

[19] Jakub Jarzyński. Analiza triad w serwisach społecznościowych. *Praca magisterska złożona na Wydziale Elektroniki i Technik Informacyjnych, Politechnika Warszawska*, 2016.

[20] Mitchell Joblin, Sven Apel, and Wolfgang Mauerer. Evolutionary trends of developer coordination: A network approach. *Empirical Software Engineering*, 22(4):2050–2094, 2017.

[21] M. P. Karpowicz, P. Arabas, and E. Niewiadomska-Szynkiewicz. Energy-aware multilevel control system for a network of Linux software routers: Design and implementation. *IEEE Systems Journal*, 12(1):571–582, 2018.

[22] Olivier Lartillot. Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *Journal Of New Music Research*, 34(4):375–393, 2005.

[23] Barbara Karolina Laskowska. Grupowanie utworów muzycznych. *Praca magisterska złożona na Wydziale Elektroniki i Technik Informacyjnych, Politechnika Warszawska*, 2019.

[24] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2006.

[25] Yang-Yu Liu and Albert-László Barabási. Control principles of complex systems. *Reviews of Modern Physics*, 88(3):035006, 2016.

[26] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[27] Stanley Milgram, Leonard Bickman, and Lawrence Berkowitz. Note on the drawing power of crowds of different size. *Journal of personality and social psychology*, 13(2):79, 1969.

[28] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[29] J.C. Nacher and T. Akutsu. On the degree distribution of projected networks mapped from bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 390(23):4636 – 4651, 2011.

[30] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[31] Ewa Niewiadomska-Szynkiewicz, Andrzej Sikora, Piotr Arabas, and Joanna Kołodziej. Control system for reducing energy consumption in backbone computer network. *Concurrency and Computation: Practice and Experience*, 25(12):1738–1754, 2013.

[32] Lawrence Page. Method for node ranking in a linked database, September 4 2001. US Patent 6,285,999.

[33] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 539–546. IEEE, 2015.

[34] Hugo Riemann. *System der musikalischen Rhythmik und Metrik*. Sändig Reprint, 1903.

[35] Souheil Ben Yacoub and Stephen Daniel James. Systems and methods for establishing ownership and delegation ownership of iot devices using domain name system services, April 3 2018. US Patent 9,935,950.

[36] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

Mariusz Kamola

Warszawa, 29 stycznia 2021 r.