VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

Jungčių prognozavimas, paremtas orientuoto tinklo klasterizacijos koeficientu

Link prediction based on directed network clustering coefficient

Magistro baigiamasis darbas

Atliko: Žilvinas Verseckas (parašas)

Darbo vadovas: prof. habil. dr. Mindaugas Bloznelis (parašas)

Recenzentas: lekt. Irmantas Radavičius (parašas)

Vilnius, 2019

Santrauka

Šiame darbe analizuojami jungties tikėtinumo tarp viršūnių porų įvertinimo metodai, kurie remiasi M. Bloznelio ir L. Leskelos (suom. Leskelä) pasiūlyta orientuoto tinklo klasterizacijos koeficiento samprata [BL16].

Šio darbo tikslas – apibrėžti jungčių tikėtinumo indeksus, besiremiančius minėtąja klasterizacijos koeficiento samprata, bei empiriškai įvertinti jungčių prognozės kokybę (naudojantis šiais indeksais) įvairiuose realiuose tinkluose.

Pirmojoje darbo dalyje suformuluojamas ir analizuojamas jungčių prognozės uždavinys, jo rezultatų vertinimo metodai bei šiuo metu praktikoje taikomi jungčių prognozavimo įrankiai. Antrojoje ir trečiojoje darbo dalyse detaliai nagrinėjama orientuoto tinklo klasterizacijos koeficiento samprata bei apibrėžiami nauji, šia samprata besiremiantys jungčių prognozės metodai: liudininkų, bendrų interesų ir įėjimo laipsnių agregacijos indeksai. Ketvirtojoje dalyje aprašomi tyrimui paruošti duomenys ir sumodeliuoti tinklai. Galiausiai, atliekami empiriniai jungčių prognozės (naudojantis naujai apibrėžtais indeksais) kokybės tyrimai.

Eksperimentų metu pastebėti aukšti kokybės metrikų įverčiai, kada jungčių prognozei yra taikomi liudininkų ir bendrų interesų indeksai. Įėjimo laipsnių agregacijos indeksas, tuo tarpu, atskleidė kitokius rezultatus: šis metodas jungtis prognozuoja atvirkščiai t.y. jungtys, kurias metodas skelbia egzistuojančiomis, iš tiesų turėtų būti skelbiamos neegzistuojančiomis ir priešingai.

Raktiniai žodžiai: jungčių prognozė, jungčių tikėtinumo indeksai, klasterizacija orientuotuose tinkluose, grafų teorija.

Summary

In this paper we analyse network link prediction methods based on a concept of digraph clustering coefficient proposed by M. Bloznelis and L. Leskelä [BL16].

The goal of this research is to define new link prediction indices derived from the clustering coefficient mentioned above. We also aim to empirically evaluate the performance of the new indices when solving link prediction problems within real-world networks.

In this paper we first define the link prediction problem and analyse possible methods of evaluating its results. We also explore and compare several already widely known link prediction methods. In the second and the third chapters we thoroughly investigate the already mentioned digraph clustering coefficient and define three new link prediction methods: the witness, the common-targets, and the indegree-aggregation indices. In the fourth chapter of this paper we describe data and networks built for our experiments. Lastly, we carry out the experiments with an aim to empirically evaluate link prediction quality of the newly defined indices.

Based on the experiments we have concluded that the witness and the common-targets indices result in high quality link prediction. In the meantime the indegree-aggregation index has shown different results - network links are predicted the opposite way. That means the links the index suggests as the existing ones are those that actually do not exist and vice versa.

Keywords: link prediction, link prediction indices, clustering in digraphs, graph theory.

Turinys

Įvadas

Realiame pasaulyje gausu sistemų, kurios susideda iš vieno ar keleto tipų elementų, tarpusavyje susietų šioms sistemoms specifiškais ryšiais. Tokios sistemos dėl savo struktūros charakteristikų ir nevisiškai žinomos prigimties įprastai vadinamos sudėtingais realaus pasaulio tinklais (angl. complex networks). Šiuose tinkluose sistemos elementus reprezentuoja tinklo viršūnės, o ryšius tarp jų – tinklo briaunos (jungtys) [AB02]. Tai iliustruoti galima socialiniais tinklais, kuriuose žmones ar kitas socialinio konteksto esybes jungia sąveikos, bendradarbiavimo ar tarpusavio įtakos ryšiai. Konkretus pavyzdys – tam tikros disciplinos mokslininkų aibė, kurioje bet kokius du mokslininkus ryšys jungia tuomet, jei šie kartu yra parašę bent vieną publikaciją; taip pat kokios nors didelės įmonės darbuotojų bendruomenė, kurios narius ryšys jungia tada, kai šie dirba prie bendro projekto.

Dėka pastaruoju metu sparčiai didėjančio kompiuterinių skaičiavimų našumo ir išaugusio visuotinio susidomėjimo didžiųjų duomenų (angl. big data) tyryba bei kompiuterių mokymusi (angl. machine learning), susidarė ypač palankios sąlygos prieiti vis daugiau detalesnės informacijos, aprašančios realaus pasaulio tinklų topologiją ir jų savybes [LK07]. To rezultatas – galimybė atlikti išsamesnius didelės apimties tinklų tyrimus, kurie šiandien jau apima aibę įvairiausio spektro disciplinų nuo informatikos, matematikos ar fizikos, iki biologijos bei socialinių mokslų [New03].

Viena iš esminių problemų, su kuria susiduriama tyrinėjant realaus pasaulio tinklus, yra tai, jog tyrimuose naudojami duomenys įprastai atspindi tik dalį realybės. Tai reiškia, kad stebimuose tinkluose tam tikros jungtys dažnai nėra užregistruotos arba jos atsiranda su laiku tinklui evoliucionuojant. Šios problemos sprendimas (dažnai dar laikomas vienu esminių tinklų analizės sprendžiamų uždavinių) yra vadinamas jungčių prognoze. Tokios prognozės tikslas yra įvertinti jungties tarp bet kokių dviejų tinklo viršūnių tikėtinumą, atsižvelgiant į turimą tinklo informaciją [LZ11].

Jungčių prognozavimas yra itin reikšmingas ir įvairiose srityse aibę taikymų turintis uždavinys. Pavyzdžiui, biologijos ar chemijos moksluose, kur turimos žinios apie tam tikras tinklais modeliuojamas sistemas dažnai yra labai ribotos, pravartu koordinuoti laboratorinius eksperimentus taip, kad būtų dirbama tik su labiausiai tikėtinomis jungtimis tinkle vietoje visų galimų jungčių nagrinėjimo. Tokiu būdu gali būti sutaupytas reikšmingas kiekis darbui reikalingų resursų [WLZ17]. Kaip kitą pavyzdį galima paminėti tinklų jungčių prognozavimą rekomendaciniais arba komerciniais tikslais. Pavyzdžiui, socialiniams tinklams, tokiems kaip *Facebook*[1], *Twitter*[2]

[1] `https://www.facebook.com`
[2] `https://twitter.com`

ar *YouTube*[3] yra svarbu savo naudotojams rekomenduoti naujus draugus ar interesus tokiu būdu siekiant pagerinti bendrą sistemos naudojimo patogumą ir kokybę bei skatinti naudotojų aktyvumą. Panašūs principai galioja ir komercinėse rekomendavimo sistemose, tokiose kaip, pavyzdžiui, taikomose *Amazon*[4], kur sistemos naudotojams prekės siūlomos pagal tai, ką šie naudotojai pirko praeityje. Be to, trūkstamų jungčių prognozavimas turi savo taikymų ir saugumo užtikrinimo srityje, pavyzdžiui, siekiant identifikuoti slaptas teroristų grupuotes [HCS+06].

Nors, kaip galima pastebėti iš pateiktų pavyzdžių, tinklo jungčių prognozavimas šiandien jau yra naudojamas įvairiose mokslo, verslo ir visuomeninėse srityse, šis uždavinys dėl savo sudėtingumo vis dar išlieka atvira problema, kuri šiuo metu yra plačiai studijuojama [WLZ17]. Kiek galima kokybiškiau jungtis prognozuojančio metodo paieška yra esminė šio **darbo motyvacija**. Prognozės kokybė šiame darbe, kaip įprasta klasifikavimo ar prognozavimo uždaviniams, yra vertinama klasikinėmis kompiuterių mokymosi ir duomenų tyrybos metrikomis, tokiomis kaip, pavyzdžiui, tikslumas, jautrumas ir specifiškumas.

Sprendžiant jungčių prognozavimo uždavinį įprastai naudojamasi dviejų tipų informacija: tinklo atributais ir/arba tinklo topologine struktūra. Pirmojo tipo informacija gali būti taikoma dirbant su daugeliu kompiuterių mokymosi algoritmų, kurių pagalba, kaip rodo tyrimai, pasiekiami itin geri prognozavimo rezultatai [HCS+06]. Kita vertus, neretai tinklo atributų informacija yra sunkiai gaunama dėl privatumo ar patikimumo problemų [WLZ17]. Tokiais atvejais pasitarnauja metodai, kurie fokusuojasi vien tik į tinklo topologinę informaciją. Šiame darbe orientuojamasi į pastarojo tipo metodus.

Jungčių prognozavimo uždavinys literatūroje dažniausiai sprendžiamas remiantis viršūnių panašumo arba briaunos tikėtinumo įvertinimu. Tai reiškia, kad kiekvienai tiesiogiai nesujungtai tinklo viršūnių porai (potencialiai jungčiai) yra priskiriamas tam tikras įvertis (požymis), dar vadinamas jungties tikėtinumo arba viršūnių panašumo indeksu. Tuomet parenkamas norimas skaičius k trūkstamų jungčių tarp didžiausią jungties tikėtinumo indeksą turinčių viršūnių porų ir šios jungtys skelbiamos prognozuojamomis jungtimis.

Šiuo metu jau egzistuoja aibė įvairių tinklo jungčių tikėtinumo įvertinimo metodų. Pavyzdžiui, preferencinio prisijungimo (angl. preferential attachment arba PA) indeksas yra apibrėžiamas kaip dviejų viršūnių laipsnių sandauga arba suma [New01]; bendrų kaimynų (angl. common neighbours arba CN) indeksas, įvertinamas kaip dviejų tinklo viršūnių bendrų kaimynų skaičius [LW71]; ar Žakardo koeficientas (angl. Jaccard arba JC), kuris yra bendrų kaimynų indekso normalizuotas variantas [Jac01]. Svarbu paminėti ir kiek naujesnius CN indekso patobulinimus, to-

[3] `https://www.youtube.com`
[4] `https://www.amazon.com`

kius kaip Adamik-Adar (angl. Adamic-Adar arba AA) [AA03] ir resursų išskyrimo (angl. resource allocation arba RA) [ZLZ09] indeksai.

Minėtieji indeksai yra nesudėtingi ir ganėtinai lengvai algoritmiškai įvertinami, tačiau tam tikrais atvejais prognozavimo rezultatai naudojantis jais gali būti nepakankamai kokybiški. Šiai problemai spręsti buvo išvystyta aibė algoritmiškai sudėtingesnių, tačiau giliau į tinklo struktūrą žvelgiančių jungčių tikėtinumo įvertinimo metodų. Dalis jų remiasi kelių paieška arba atsitiktiniu klaidžiojimu tinkle [ZRM+09], dalis – tam tikromis svarbiomis realaus pasaulio tinkluose stebimomis topologinėmis charakteristikomis, tokiomis kaip klasterizacija [WMR+15].

Daugelyje realaus pasaulio tinklų, kaip rodo tyrimai, stebima viršūnių tendencija klasterizuotis. Tai reiškia, jog tinklo viršūnės pasižymi savybe burtis į tankias grupes retame, globaliu požiūriu, tinkle (angl. sparse network) [BL16]. Tokią tinklų savybę kiekybiškai įvertina klasterizacijos koeficientas. Neorientuotame tinkle lokalus viršūnės klasterizacijos koeficientas gali būti suprantamas kaip per šią viršūnę einančių trikampių (ilgio trys ciklai) skaičiaus santykis su šios viršūnės kaimynų porų skaičiumi. Šį dydį galima interpretuoti ir kaip tikimybę, kad viršūnės kaimynai tarpusavyje yra taip pat kaimynai [New03].

Kalbant apie orientuotus tinklus, klasterizacijos samprata juose gali būti suvokiama aibe įvairių būdų [BL16]. Vienas tokių būdų šiame darbe nagrinėjamas kaip pagrindas orientuoto tinklo jungčių prognozei. Orientuoto tinklo jungčių prognozės metodų paiešką iš esmės **motyvuoja** tai, kad, nepaisant didelio tokių tinklų paplitimo, didžioji dalis literatūroje sutinkamų jungčių prognozės metodų yra taikytini tik neorientuotiems tinklams.

Šiame darbe nagrinėjamas jungties tikėtinumo tarp viršūnių porų įvertinimo metodas, kuris remiasi M. Bloznelio ir L. Leskelos (suom. Leskelä) apibrėžiama, viena naujesnių, jungčių prognozei dar neišmėgintų orientuoto tinklo klasterizacijos sampratų: *viršūnės sekėjai, tikėtina, turi dar ir kitų bendrai sekamų viršūnių* [BL16]. Tokio tipo klasterizacijos koeficientas tinkluose įvertina tikimybę, kad dvi viršūnės, turinčios orientuotą briauną į bendrą viršūnę, turės bendrų jungčių į dar bent vieną kitą tinklo viršūnę [BL16]. Toks tinklo motyvas, kaip pastebima, dominuoja daugybėje realaus pasaulio tinklų, tokių kaip, pavyzdžiui, citavimo ar internetiniai socialiniai tinklai [ZLW+13]. Minėtojo klasterizacijos koeficiento samprata darbe taip pat išplečiama keliomis jos variacijomis, į koeficiento skaičiavimą įtraukiančiomis papildomus faktorius, tokius kaip viršūnių laipsniai bei bendrai sekamų viršūnių skaičius.

Šio darbo **tikslas** – apibrėžti jungčių tikėtinumo indeksus, besiremiančius orientuoto tinklo klasterizacijos koeficiento samprata [BL16], bei empiriškai įvertinti jungčių prognozės kokybę (naudojantis šiais indeksais) įvairiuose tinkluose. Tikslui pasiekti iškeliami penki uždaviniai:

1. Išanalizuoti šiandien jau praktikoje taikomus jungčių prognozavimo įrankius bei metodus.
2. Išanalizuoti orientuoto tinklo klasterizacijos koeficiento sampratą ir jos praplėtimus [BL16].
3. Apibrėžti ir išnagrinėti jungčių tikėtinumo indeksus, paremtus orientuoto tinklo klasterizacijos koeficientu ir jo praplėtimais.
4. Surinkti tyrimo duomenis bei sumodeliuoti kelis internetinius socialinius tinklus.
5. Empiriškai ištirti jungčių prognozės kokybę realiuose tinkluose, naudojantis darbe apibrėžtais jungčių tikėtinumo indeksais.

Šio darbo **rezultatai** yra trys nauji orientuoto tinklo klasterizacijos samprata [BL16] besiremiantys jungčių tikėtinumo indeksai su empiriniu jungčių prognozės kokybės įvertinimu (naudojantis šiais indeksais). Be to, vienas iš darbo rezultatų yra ir realių tinklų, taikytų jungčių prognozės kokybės tyrimuose, surinkimas, sumodeliavimas ir ištyrimas.

Darbe nagrinėjamų jungčių tikėtinumo indeksų **tyrimas** remiasi keleto skirtingo dydžio orientuotų tinklų empirine analize. Kiekvieno tiriamo tinklo duomenys dalinami į nepersikertančias mokymo ir testavimo aibes. Pastarojoje dalis originalaus tinklo jungčių yra paslepiamos ir naudojamos kaip testavimo jungtys. Svarbu pabrėžti, kad siekiant išlaikyti tinklo vidinę struktūrą paslepiamas tik santykinai nedidelis jungčių skaičius [WLZ17]. Galiausiai, įvertinus visų trūkstamų jungčių tikėtinumo įverčius, prognozuojamos jungtys. Prognozės kokybė matuojama standartinėmis klasifikavimo ir prognozės uždavinių metrikomis.

Darbo **struktūra** tiesiogiai atspindi iškeltus uždavinius ir yra skirstoma į penkias esmines dalis. Pirmojoje dalyje suformuluojamas ir analizuojamas jungčių prognozės uždavinys, jo rezultatų vertinimo metodai bei šiuo metu praktikoje taikomi jungčių prognozavimo įrankiai (pirmasis uždavinys). Antrojoje dalyje nagrinėjama orientuoto tinklo klasterizacijos koeficiento samprata (antrasis uždavinys). Trečiojoje darbo dalyje apibrėžiami ir analizuojami orientuoto tinklo klasterizacijos samprata besiremiantys jungčių prognozės metodai (trečiasis uždavinys). Ketvirtojoje dalyje aprašomi tyrimui paruošti duomenys ir sumodeliuoti tinklai (ketvirtasis uždavinys). Galiausiai, atliekami empiriniai jungčių prognozės, paremtos orientuoto tinklo klasterizacijos samprata, kokybės tyrimai (penktasis uždavinys).

1. Tinklo jungčių prognozė

Dėl didelio paplitimo ir vis augančio aktualumo sudėtingų realaus pasaulio tinklų tyrimai tapo ne tik itin svarbia tinklų teorijos dalimi bet ir vaidina vis reikšmingesnį vaidmenį aibėje kitų mokslo sričių, tokių kaip biologija, fizika, socialiniai, kompiuterių ar duomenų tyrybos mokslai. Įvairių disciplinų atstovai bei įvairaus tipo organizacijos deda didžiules pastangas siekdami suprasti sudėtingų realaus pasaulio tinklų raidą, savybes, šių savybių įtaką tinklų vidinei struktūrai bei jungčių tarp tinklo elementų formavimosi tendencijas.

Labai svarbus sudėtingų realių tinklų tyrybos uždavinys yra tinklo jungčių prognozė, kurios tikslas – įvertinti, kiek tikėtina yra nesanti tiesioginė jungtis tarp dviejų tinklo viršūnių. Ši tikimybė yra įvertinama remiantis turima tiriamo tinklo informacija, tokia kaip jame jau esančios jungtys, šių jungčių ar tinklo viršūnių atributai, tinklo raidos istoriniai duomenys [LZ11].

Jungčių prognozė šiandien turi aibę taikymų įvairiausiose mokslo srityse. Pavyzdžiui, biologiniuose baltymų sąveikų (angl. protein interaction), medžiagų apykaitos (angl. metabolic) ar mitybos tinkluose. Tokios sistemos kaip baltymų sąveikų tinklai yra ypač didelės apimties ir milžiniška dalis jungčių jose yra nežinoma (pavyzdžiui, mes neturime informacijos apie maždaug 98 % molekulių sąveikų žmogaus ląstelėse). Jungčių prognozė čia leidžia identifikuoti nežinomas jungtis ir dirbti tik su labiausiai tikėtinomis, kas padeda sutaupyti didelį kiekį materialinių ir laiko sąnaudų, kurias tektų skirti jungčių identifikavimui laboratoriniais eksperimentais ar „aklam“ visų galimų atvejų analizavimui [WLZ17].

Socialiniuose tinkluose jungčių prognozavimas taip pat plačiai taikomas trūkstamų ryšių tarp tinklo elementų identifikavimui. Be to šis metodas pasitarnauja ir klaidingoms jungtims tinkle identifikuoti [LZ11], kurios konstruojant tokius tinklus neretai atsiranda dėl socialinių faktorių, tokių kaip žmonių šališkumas ar skirtingas dalykų interpretavimas.

Be trūkstamų jungčių prognozavimo, socialinių tinklų tyryboje ypač svarbus yra ir būsimų jungčių numatymas [LZ11]. Pavyzdžiui, labai tikėtinos, tačiau dar neegzistuojančios pažintys gali būti rekomenduojamos dviems žmonėms socialiniame tinkle. Tai padeda kurti naujas pažintis, o, tarkime, pažinčių portalo savininkams tai neša pelną ir didina naudotojų lojalumą. Kaip kitas pavyzdys gali būti pateikiamas prekių ar paslaugų rekomendavimas komerciniais tikslais. Šiuo atveju tam tikros prekės ar paslaugos gali būti rekomenduojamos vartotojams, remiantis jų pirkimų istorija (t.y. ryšiais su kitomis prekėmis ar paslaugomis), vartotojų sąryšiais su kitais vartotojais ar tam tikrais žinomais vartotojo atributais.

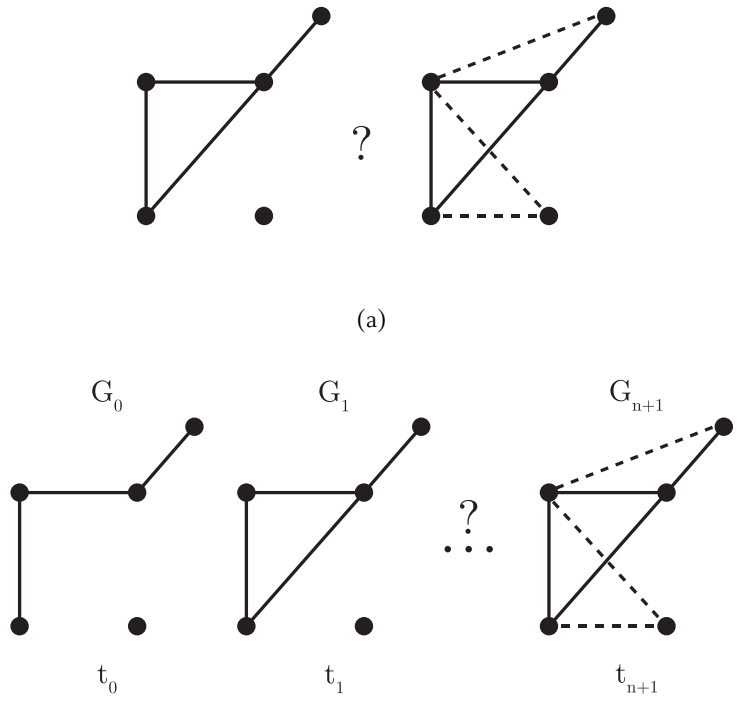
Taigi, iš esmės jungčių prognozės uždavinius galima padalinti į du pagrindinius tipus: *struktūrinę* ir *laikinę* prognozes [Puj15] (žr. 1 pav.).

Struktūrinė jungčių prognozė

Šių uždavinių tikslas yra nustatyti trūkstamas ar neužregistruotas tinklo jungtis, kurios, veikiausiai, egzistuoja realiame tinkle. Šiuo atveju jungties tikėtinumui nustatyti daugiausiai remiamasi žinoma tinklo struktūra (pavyzdžiui, žinomomis jungtimis). Toks prognozavimas yra ypač plačiai taikomas genetiniuose, baltymų sąveikų tinkluose ar medicinoje [Puj15]. Taip pat ir visuomenės saugumo srityje siekiant identifikuoti galimai egzistuojančius ryšius tarp nusikaltėlių ar teroristinių grupuočių [HCS+06].

Laikinė jungčių prognozė

Šie uždaviniai orientuojasi į naujų galimų jungčių prognozavimą, remiantis tiriamo tinklo raida bėgant laikui. Tai reiškia, kad prognozės metu nagrinėjama tinklo struktūra laiko momentu t_0, o siekiamas rezultatas – kokybiškas naujų jungčių numatymas laiko momentu $t_0 + \Delta t$ [Puj15]. Tokio tipo uždaviniai ypač dažnai sprendžiami rekomendacinėse sistemose, plačiai taikomose elektroninės komercijos portaluose, paieškos sistemose, siekiant naudotojui rodyti tik jam aktualiausią informaciją, taip pat ir internetiniuose socialiniuose tinkluose, rekomenduojant draugus ar interesų grupes (pavyzdžiai: *Facebook*, *Twitter*, *YouTube*). Dar vienas svarbus tokių uždavinių taikymas yra ateities bendradarbiavimo prognozė, pavyzdžiui, mokslininkų akademinio bendradarbiavimo socialiniuose tinkluose [Puj15].



1 pav. Jungčių prognozavimo uždavinių tipai. (a) – struktūrinė prognozė trūkstamoms jungtims identifikuoti, (b) – laikinė prognozė būsimoms jungtims numatyti. Parengta pagal: [Puj15].

Likusioje šio skyriaus dalyje pirmiausiai formaliai apibrėžiamas jungčių prognozės uždavinys, vėliau – tolesniame darbe naudojami žymėjimai, terminologija bei prognozės metodų vertinimo kriterijai.

1.1. Uždavinio apibrėžimas ir terminologija

Jungčių prognozės uždavinys, kaip jau minima anksčiau, siekia nustatyti tinkle neužregistruotas arba ateityje susiformuosiančias jungtis. Šiam uždaviniui spręsti yra taikoma aibė įvairiausių metodų, kurie remiasi tinklo komponentų požymiais, tinklo struktūra, topologija ar kitomis jo savybėmis [Puj15].

Jungčių prognozavimo uždavinį bendrai galima formuluoti taip: tarkime, turime tinklą $G(V,E)$, kur V yra šio tinklo viršūnių aibė, o E – jo briaunų aibė. Visas tokiame tinkle įmanomas jungtis pažymėjus aibe U, visas trūkstamas jungtis jame atspindės aibių skirtumas $U \setminus E$. Tuomet daroma prielaida, kad aibėje $U \setminus E$ vis tik egzistuoja tam tikros jungtys $(u,v) \notin E, u,v \in V$, kurios arba nėra užregistruotos, bet egzistuoja realiame tinkle, arba atsiras šiam tinklui evoliucionuojant [LZ11]. Jungčių prognozės tinklas – identifikuoti šias jungtis.

Naujų jungčių prognozavimo uždavinys tinkui evoliucionuojant (laikinė prognozė) gali būti aprašomas šiek tiek patikslinus aukščiau pateikiamą bendrąjį uždavinį. Šiuo atveju siekiama nuspėti naujas jungtis tinkle tam tikru jo raidos momentu t_{n+1}, kai yra žinoma, kaip šis tinklas evoliucionavo iki tam tikro momento t_n. Taigi, visas tinklas gali būti apibūdintas grafų seka $G = \langle G_0, G_1, \ldots, G_n \rangle$ skirtingais jo raidos momentais $\langle t_0, t_1, \ldots, t_n \rangle$. Kiekvienas sekos G grafas G_i, turi savo viršūnių ir briaunų aibes V_i ir E_i atitinkamai. Ši grafų seka G taip pat gali būti suprantama ir kaip grafas $G(V,E)$, kur $V = \bigcup_{i=0}^{n} V_i$, o $E = \bigcup_{i=0}^{n} E_i$. Naujų jungčių prognozavimo uždavinio tikslas, panašiai kaip ir bendruoju atveju, yra identifikuoti potencialias naujas jungtis tiriamo tinklo raidos momentu t_{n+1} arba $t_{n+k}, k \in \mathbb{N}$ [Puj15].

1.2. Prognozės metodų vertinimas ir metrikos

Didelė dalis tradicinių jungčių prognozės metodų nagrinėjamą uždavinį sprendžia potencialioms jungtims priskirdami teigiamą arba neigiamą požymį (klasę), atitinkamai nurodantį, ar jungtis prognozuojama kaip egzistuojanti, ar ne. Kitas plačiai taikomas būdas yra kiekybiškai įvertinti, kiek tikėtina yra konkreti jungtis, ir grąžinti tikėtinumo įverčių mažėjimo tvarka surikiuotą potencialių jungčių sąrašą (su šiais įverčiais) [ZYW14]. Tuomet tam tikras kiekis didžiausius tikėtinumo įverčius turinčių potencialių jungčių gali būti skelbiamas kaip trūkstamos arba susiformuosiančios ateityje jungtys. Taigi, jungčių prognozės metodų rezultatai įprastai būna arba sąrašas teigiamai

ir neigiamai suklasifikuotų jungčių, arba sąrašas jungčių su joms priskirtais tikėtinumo įverčiais. Šie rezultatai tuomet gali būti naudojami prognozei taikyto metodo įvertinimui.

Jungčių prognozės metodų įvertinimas yra taip pat nemažai iššūkių kelianti užduotis. Viena iš to priežasčių yra faktas, kad didžioji dalis realių tinklų yra reti (angl. sparse). Tai reiškia, kad tik labai nedidelė visų galimų jungčių dalis iš tiesų egzistuoja tinkle. Remiantis tuo galima teigti ir tai, kad naujų ar trūkstamų jungčių tinkle būna taip pat labai mažai, palyginus su visomis potencialiomis jungtimis, kurios galbūt niekada ir nesusiformuos.

Vienas iš jau minėtų prognozės būdų yra jungčių prognozės rezultatu skelbti tam tikrą kiekį $k \in \mathbb{N}$ didžiausius įverčius turinčių jungčių. Tokiu atveju dažnai nėra aišku, kokiais kriterijais reikėtų vadovautis parenkant slenkstį k. Panašus metodas yra ir tam tikro slenksčio T pasirinkimas, kuriuo remiantis, visos potencialios jungtys su tikėtinumo įverčiais didesniais arba lygiais T būtų skelbiamos prognozuojamomis jungtimis. Šiuo atveju iškyla pirmam atvejui analogiška problema, kadangi įprastai nėra aišku, kokią slenksčio reikšmę prasminga pasirinkti. Dažnai literatūroje sutinkamas sprendimas šiai problemai yra prognozės metodo vertinimas keičiant slenksčių T arba k reikšmes. Tokiu atveju šios reikšmės kinta tam tikrame intervale ir kiekvienai galimai šių slenksčių reikšmei yra įvertinama prognozės metodo kokybė. Tokio jungčių prognozės vertinimo metodo rezultatas yra sąrašas tam tikrų kokybės įverčių, priklausančių nuo slenksčių T arba k reikšmių.

Standartiškai tinklo jungčių prognozės metodo rezultatus galima vertinti pasitelkiant kompiuterių mokymosi priemones, kur teisingai teigiamas (angl. true positive) atvejis reiškia, kad prognozuojama jungtis iš tikrųjų egzistuoja, klaidingai teigiamas (angl. false positive) – teigiamas jungties egzistavimas yra prognozuojamas klaidingai ir analogiškai neigiamos prognozės atveju (žr. 1 lentelę).

1 lentelė. Klasifikavimo matrica (angl. classification/confusion matrix).

	Teigiama prognozė	**Neigiama prognozė**
Iš tikrųjų teigiama	Teisingai teigiamas (TT)	Klaidingai neigiamas (KN)
Iš tikrųjų neigiama	Klaidingai teigiamas (KT)	Teisingai neigiamas (TN)

Deja bet įprastai nėra žinoma, kurios jungtys tinkle yra neužregistruotos arba kurios jų atsiras ateityje (kitaip nereikėtų jungčių prognozės). Dėl šios priežasties prognozės metodo efektyvumas dažnai vertinamas žinomų jungčių aibę E dalinant į dvi dalis: mokymo aibę E_M, kuri yra naudojama kaip žinoma informacija apie tinklo jungtis, ir testavimo aibę E_T, kuri yra naudojama metodo prognozės rezultatų įvertinimui [LZ11]. Svarbu pabrėžti, kad minėtosios aibės negali persidengti: $E_M \cup E_T = E$ ir $E_M \cap E_T = \varnothing$. Aibės E dalinimas į mokymo ir testavimo aibes gali

būti paremtas tinklo evoliucija arba būti atsitiktinis. Pirmuoju atveju, tam tikra dalis naujausių jungčių (jei žinomas jungties susiformavimo laikas) gali būti parenkama kaip testavimo aibė E_T, o likusios jungtys – kaip mokymo aibė E_M. Atsitiktinio dalinimo pagrindinis trūkumas yra tai, kad nuo laiko priklausomuose tinkluose gali būti prarandama svarbi tinklo topologijos informacija. Be to, dalis jungčių gali niekada nepatekti į testavimo aibę, kai dalis gali į ją patekti kelis kartus. Tai yra potenciali vieta išaugti statistiniam šališkumui (angl. statistical bias). Toks apribojimas gali būti pašalinamas naudojant *k-blokų kryžminę patikrą* (angl. *k-fold cross-validation*), kurios pagalba žinomų jungčių aibė E yra atsitiktinai padalinama į k blokų ir kiekvienoje testavimo iteracijoje vienas blokas yra pasirenkamas kaip testavimo aibė E_T, o likusieji $k-1$ blokų – kaip mokymo aibė E_M. Kryžminės patikros metodas tuomet kartojamas k kartų kiekvieną bloką panaudojant kaip testavimo aibę tiksliai po vieną kartą. Tokiu būdu visos žinomos jungtys yra panaudojamos tiek kaip mokymo, tiek ir kaip testavimo subjektai [LZ11].

Šiandien egzistuoja aibė metrikų, kurios gali būti naudojamos įvertinti prognozei taikomo metodo efektyvumą ir prognozės rezultatų kokybę. Šios metrikos turėtų būti apskaičiuojamos išmėginus testavimo aibę su jungčių prognozės metodo rezultatais mokymo aibėje. Žemiau pateikiamas sąrašas tokių metrikų [LZ11; Puj15; ZYW14] (metrikos žemiau remiasi klasifikavimo matricos terminologija (žr. 1 lentelę)).

- *Tikslumas* (angl. accuracy). Tikslumas atspindi, kokia dalis visų prognozuojamų jungčių yra prognozuojama teisingai.

$$\textit{Tikslumas} = \frac{TT + TN}{TT + TN + KT + KN} \tag{1}$$

- *Jautrumas* (angl. precision). Jautrumas nurodo, kokią dalį visų teigiamai prognozuojamų jungčių sudaro teisingai teigiamai prognozuojamos jungtys.

$$\textit{Jautrumas} = \frac{TT}{TT + KT} \tag{2}$$

  Tuo atveju, kada jungčių prognozės metodas parenka $k \in \mathbb{N}$ labiausiai tikėtinų jungčių (pvz. metodai priskiriantys jungčių tikėtinumo įverčius), jautrumo lygtis gali būti užrašoma paprasčiau (žr. (3) lygybę).

$$\textit{Jautrumas} = \frac{TT}{k} \tag{3}$$

- *Specifiškumas* (angl. recall). Specifiškumas atspindi, kokią dalį visų iš tiesų teigiamų atvejų sudaro teisingai teigiamai prognozuojamos jungtys.

$$\textit{Specifiškumas} = \frac{TT}{TT + KN} \tag{4}$$

- F_1 *indeksas*. F_1 indeksas yra harmoninis jautrumo ir specifiškumo vidurkis.

$$F_1 = \frac{2 \cdot \textit{Jautrumas} \cdot \textit{Specifiškumas}}{\textit{Jautrumas} + \textit{Specifiškumas}} \tag{5}$$

- *ROC kreivė* (angl. Receiver Operating Characteristic). ROC kreivės sugeneruojamos horizontalioje ašyje atvaizduojant klaidingai teigiamų prognozių rodiklį $KTR = KT/(KT + TN)$ (angl. false positive rate), o vertikalioje – teisingai teigiamų $TTR = TT/(TT + KN)$ (angl. true positive rate). Kreivė sudaroma visoms slenksčio tarp minimalios ir maksimalios šių dviejų dydžių reikšmėms (tarp nulio ir vieneto). Idealios jungčių prognozės atveju, teisingai teigiamų prognozių rodiklis yra lygus vienetui (visos teigiamai prognozuojamos yra parenkamos teisingai). Neigiamų prognozių dalis atitinkamai yra lygi nuliui (visos neigiamai prognozuojamos jungtys yra parenkamos teisingai). Tai būtų ekvivalentu taškui $(0{,}1)$ ROC kreivėje (žr. 3a pav.). Pats blogiausias galimas jungčių prognozės įrankis yra atsitiktinis prognozuojamų jungčių parinkimas. Tokiu atveju ROC kreivė yra artima įstrižainei iš taško $(0{,}0)$ į tašką $(1{,}1)$.

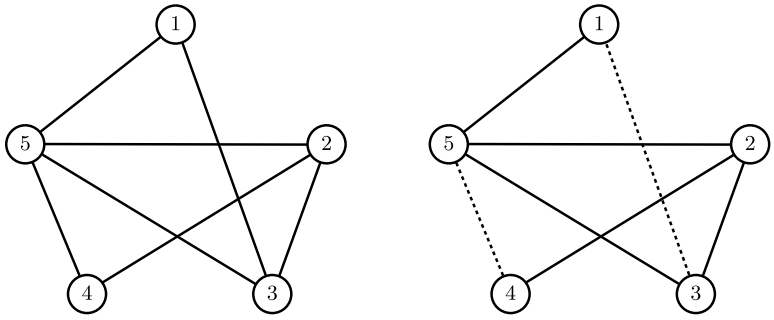
- *AUROC* (angl. Area Under ROC Curve). AUROC galima įsivaizduoti kaip plotą po ROC kreive. Šis dydis atitinka tikimybę, kad atsitiktinai pasirinktai trūkstamai jungčiai (pvz. iš testavimo aibės E_T) prognozės metodas priskiria didesnį tikėtinumo įvertį, nei atsitiktinai pasirinktai neegzistuojančiai jungčiai. Šį dydį galima įvertinti dviem esminiais žingsniais. Pirmiausia, visoms potencialioms (trūkstamoms ir neegzistuojančioms) jungtims priskiriamas tikėtinumo įvertis. Tada atsitiktinai ir nepriklausomai atliekama n palyginimų tarp bet kokių trūkstamos ir neegzistuojančios jungčių porų. Jei tarp šių palyginimų rezultatų n' kartų trūkstama jungtis turi didesnį tikėtinumo įvertį nei neegzistuojanti jungtis ir n'' kartų tokį patį, tada AUROC yra:

$$\text{AUROC} = \frac{n' + 0{,}5n''}{n}. \tag{6}$$

Didesnės AUROC reikšmės indikuoja geresnius prognozės rezultatus. Reikšmė $0{,}5$ reprezentuoja visišką atsitiktinumą, todėl tinklo jungčių prognozės metodai gali būti lyginami

tarpusavyje pagal tai, kiek daug jie perkopia šią ribą [LZ11]. Empiriškai apytikslę dydžio AUROC reikšmę galima apskaičiuoti kaip skaitinį integralą (plotą po ROC kreive).

ROC kreivė su AUROC yra, ko gero, vieni populiariausių jungčių prognozės įvertinimo įrankių ir dažnai gerai atspindi tiriamo metodo prognozės kokybę. Kita vertus, jei mokymo aibė yra labai didelė arba joje stebimas itin didelis klasių (jungtis egzistuoja ir jungtis neegzistuoja) disbalansas, ROC kreivė gali klaidinti. Tarkime, atliekamas jungčių prognozės metodo įvertinimas retame tinkle (didelis klasių disbalansas). Tokiame tinkle jungčių prognozės metodas gali su didele tikimybe teisingai nuspėti didelę dalį trūkstamų jungčių, tačiau šis metodas taip pat parinks ir labai daug neigiamai teigiamų atvejų (daug jungčių bus skelbiamos kaip egzistuojančios klaidingai). Nors toks rezultatas indikuoja prastą prognozės kokybę, ROC kreivė ir atitinkamai AUROC indikuos gerus rezultatus (dėl didžiulės neigiamos klasės persvaros) [LLC10].



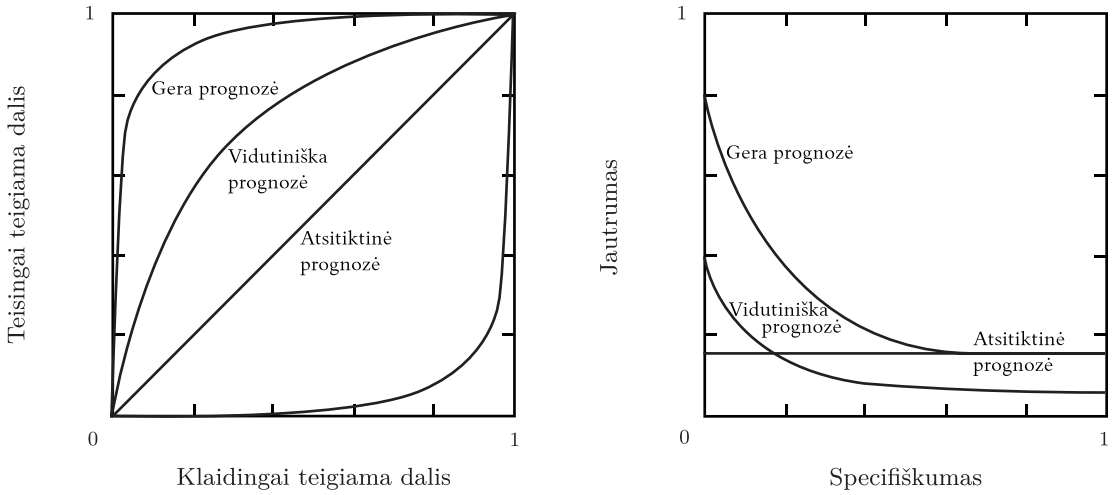
(a) Tiriamas grafas (b) Mokymo pografis

2 pav. AUROC metrikos įvertinimo pavyzdys. Šiame pavyzdiniame grafe yra penkios viršūnės ir septynios egzistuojančios bei trys neegzistuojančios briaunos $((1,2),(1,4)$ ir $(3,4))$. Prognozės metodui įvertinti iš pradinio grafo reikia pasigaminti mokymo ir testavimo pografius. Pasirenkame briaunas $(1,3)$ ir $(4,5)$ kaip testavimo pografį ir paslepiame jas (pažymime punktyru) taip pradinį grafą paversdami mokymo pografiu. Tarkime, tam tikras jungčių prognozės metodas visoms nematomoms mokymo grafo jungtims priskyrė tam tikrus tikėtinumo įverčius: $s_{12} = 0{,}4$, $s_{13} = 0{,}5$, $s_{14} = 0{,}6$, $s_{34} = 0{,}5$ ir $s_{45} = 0{,}6$. Kad apskaičiuotume AUROC, palyginame testavimo pografio egzistuojančių briaunų tikėtinumo įverčius su neegzistuojančių briaunų tikėtinumo įverčiais mokymo grafe: $s_{13} > s_{12}, s_{13} < s_{14}, s_{13} = s_{34}, s_{45} > s_{12}, s_{4,5} = s_{3,4}$ ir $s_{45} > s_{34}$. Todėl AUROC $= (3 + 2 \cdot 0{,}5)/6 \approx 0{,}67$. Parengta pagal: [LZ11].

- *Jautrumo-specifiškumo kreivė* (angl. Precesion-Recall Curve). Jautrumo-specifiškumo kreivėse kiekvienas taškas vaizduoja jautrumo ir specifiškumo porą kintant slenksčio reikšmei tarp maksimalaus ir minimalaus šių dviejų dydžių įverčių (panašiai kaip ROC kreivėse). Horizontali ašis čia yra skiriama specifiškumui, o vertikali – jautrumui. Plotas po šio tipo kreivėmis (žinomas kaip AUPR) taip pat gali būti naudojamas panašiu tikslu kaip ir AUROC atveju (žr. 3b pav.). Bazinė linija jautrumo-specifiškumo kreivėje simbolizuoja

atsitiktinai jungtis prognozuojančio metodo prognozės kokybę. Taigi, kuo tiriamo metodo jautrumo-specifiškumo kreivė yra labiau nutolusi virš bazinės linijos, tuo geresnė yra jungčių prognozės metodo rezultatų kokybė. Panašiai kaip ir ROC kreivės atveju, idealios prognozės atveju jautrumo ir specifiškumo įverčiai yra vienetai.

Svarbu atkreipti dėmesį, kad jautrumo-specifiškumo kreivės gali būti ypač naudingos tyrinėjant duomenis su labai nevienodu klasės pasiskirstymu (ignoruoja teisingai neigiamus atvejus) [Puj15], kuris ypač dažnai pasitaiko sprendžiant jungčių prognozės uždavinius (įprastai naujų ar trūkstamų jungčių būna labai mažai palyginus su visomis galimomis jungtimis, kurios galbūt niekada nesusiformuos).



(a) ROC kreivės

(b) Jautrumo-specifiškumo kreivės

3 pav. ROC ir jautrumo-specifiškumo kreivių interpretavimo pavyzdžiai. Kuo arčiau ROC kreivė artėja prie taško $(0,1)$ (ideali prognozė), tuo geresnę tiriamo metodo prognozės kokybę ROC indikuoja. Atitinkamai, kuo toliau, jautrumo-specifiškumo kreivė yra nutolusi nuo bazinės linijos (atsitiktinės prognozės) ir kuo arčiau taško $(0,1)$, tuo geresnę tiriamo metodo prognozės kokybę ši kreivė indikuoja. Apversta ROC kreivė parodo, kad jungties egzistavimo ir neegzistavimo požymiai yra parenkami priešingai (t.y. prognozuojamomis jungtimis reikėtų skelbti tas jungtis, kurias metodas parenka kaip neegzistuojančias).

Apibendrindami galime pastebėti, kad jungčių prognozės metodų, skaičiuojančių jungčių tikėtinumą, rezultatams vertinti yra taikytinos visos aukščiau nagrinėjamos metrikos, be to, ROC ir jautrumo-specifiškumo kreivės bei AUROC ir AUPR gali pateikti ir detalesnių įžvalgų. Tuo tarpu, jungčių prognozės metodai, skirstantys jungtis į dvi grupes (priskirdami teigiamą arba neigiamą jungties egzistavimo požymį) įprastai gali būti vertinami tik tikslumo, jautrumo , specifiškumo ir jų išvestinėmis metrikomis. Taip yra todėl, kad pastarojo tipo jungčių prognozės metodų rezultatai neturi pareinamų reikšmių (tik teigiamas/neigiamas, egzistuoja/neegzistuoja).

1.3. Tinklo jungčių prognozės metodai

Šiandien yra žinoma aibė tinklo jungčių prognozės metodų. Vieni jų naudojasi tinklo viršūnių atributais, kiti – tik tinklo topologine (struktūrine) informacija, treti, dar vadinami hibridiniais metodais, – remiasi tiek tinklo topologija, tiek viršūnių informacija [Puj15].

Naudojantis viršūnių atributais besiremiančiais jungčių prognozės metodais, įprastai atsižvelgiama į tam tikrą papildomą, to tinklo elementams būdingą informaciją. Kaip pavyzdį čia galima paminėti mokslinių publikacijų citavimo tinklą, kurį sudaro moksliniai straipsniai ir juos jungiantys citavimo ryšiai. Greta šios struktūrinės minėtojo tinklo informacijos yra žinomi ir papildomi jo viršūnėms būdingi atributai, tokie kaip straipsnio turinys, jo autoriai, mokslinė sritis ir panašiai. Ši informacija neretai gali būti itin naudinga, pavyzdžiui, ieškant panašumų tarp nesujungtų tinklo viršūnių porų. Tuomet, remiantis šiais panašumais, gali būti prognozuojamos jungtys tinkle.

Tinklo topologija besiremiantys prognozės metodai dirba tik su tinklo struktūrine informacija be jokių papildomų duomenų apie to tinklo elementus. Šie metodai nagrinėja, pavyzdžiui, kokiu dėsningumu tinkle susiformavimo jame jau esančios jungtys ir kaip jos kinta bėgant laikui. Remiantis tokio tipo informacija tuomet bandoma įvertinti, kiek tikėtina yra tam tikra trūkstama ar būsima jungtis [Puj15]. Šiam tikslui gali būti pasitelkiamos ir tam tikros plačiai žinomos tinklų topologinės metrikos, tokios kaip klasterizacijos koeficientas ar vidutinis kelio ilgis tarp tinklo viršūnių [Hua06].

Reikėtų pastebėti, kad tinklo elementų atributais besiremiantys prognozės metodai neretai gali būti naudingi dirbant su labai fragmentuotais (susidedančiais iš daug nejungių komponentų) arba kitaip topologiškai neinformatyviais tinklais. Tuo tarpu į tinklo struktūrą orientuoti metodai yra itin efektyvūs tada, kada tinklo atributų informacija yra sunkiai prieinama arba iš viso nežinoma. Šiuos metodus taip pat galima laikyti bendresniais ir mažiau priklausomais nuo konkrečių tinklų specifikos. Kartais gali būti paranku ir kombinuoti abiejų minėtų tipų metodus išvedant hibridinius jungčių prognozės modelius [Puj15].

Kadangi šiame darbe nagrinėjama tam tikra tinklo klasterizacijos koeficiento samprata ir jos taikymas jungtims prognozuoti, toliau daugiau orientuosimės į tinklo topologija paremtus tinklo jungčių prognozės metodus.

Šiandien egzistuoja ne vienas tinklo topologija besiremiantis jungčių prognozės metodas, taikytinas trūkstamų ar būsimų jungčių tinkle tikėtinumui įvertinti. Šie metodai, naudodamiesi tam tikrais tinklo struktūros atributais, kiekvienai tiesiogiai nesujungtų tinklo viršūnių porai (x,y) (potencialiai jungčiai) priskiria tam tikrą skaitinį įvertį s_{xy}, dažnai dar vadinamą viršūnių panašumo arba jungties tikėtinumo indeksu [Puj15]. Formaliai aibę visų tiesiogiai nesujungtų viršūnių porų

tinkle $G(V,E)$ žymėsime $L = \{(x,y)\}$, kur $x,y \in V, x \neq y$, o $(x,y) \notin E$.

Naudojantis topologija grįstais metodais jungčių prognozė gali būti organizuojama kiekvienai viršūnių porai $(x,y) \in L$ priskyrus dydį s_{xy} ir atrinkus tam tikrą kiekį didžiausius s_{xy} įverčius turinčių viršūnių porų. Šios poros skelbiamos prognozės rezultatais.

Tinklo topologija paremti jungčių prognozės metodai gali būti skirstomi į kategorijas pagal tam tikras šių metodų ypatybes. Pavyzdžiui, į laikinius ir statinius, priklausomai nuo tuo, ar metodas atsižvelgia į nagrinėjamo tinklo dinamiką ir raidą [Puj15]. Taip pat jie gali būti skirstomi į lokalius ir globalius, pagal tai, kokiu lygiu yra skaičiuojami jungčių tikėtinumo įverčiai. Kai šie skaičiavimai yra atliekami, pavyzdžiui, viršūnių porai, prognozės metodas priskiriamas lokaliems metodams, kada tai daroma visam tinklui arba potinkliui – globaliems [LZ11].

Toliau šiame skyriuje nagrinėjami jungčių prognozės metodai, pagal tikėtinumo įverčio s_{xy} skaičiavimo lygį, skirstomi į dvi grupes: lokalius ir globalius.

1.3.1. Lokalūs tinklo jungčių prognozės metodai

Lokalus jungties tikėtinumo įvertis s_{xy} gali būti skaičiuojamas įvairiais būdais, naudojantis aibe skirtingų tinklo topologijos atributų. Šie atributai gali būti paremti viršūnių kaimynyste, tam tikromis viršūnės savybėmis ar tam tikra šių savybių kombinacija (agregacija).

1.3.1.1. Viršūnių kaimynyste paremti metodai

Viršūnių kaimynyste grįsti metodai remiasi idėja, kad jungties tarp tinklo viršūnių tikimybė išauga tada, kada šios viršūnės tarpusavyje dalinasi tais pačiais arba kokiu nors aspektu panašiais kaimynais [Puj15]. Viršūnėms x ir y bendrais kaimynais yra laikomos tos viršūnės, kurios yra sujungtos briauna tiek su viršūne x, tiek ir su viršūne y. Šių viršūnių kaimynų aibes žymėsime $\Gamma(x)$ ir $\Gamma(y)$.

Čia egzistuoja aibė būdų įvertinti jungties $(x,y) \in L$ tikėtinumo indeksą s_{xy} remiantis viršūnių kaimynystės samprata. Keletas dažniausiai sutinkamų ir plačiau taikomų metodų pateikiami žemiau [HCS+06; LZ11; Puj15].

- *Bendrų kaimynų* indeksas – vienas paprasčiausių jungties tikėtinumo įverčių, reprezentuojamas paprasčiausiai dviejų viršūnių bendrų kaimynų skaičiumi. Šis metodas iš esmės remiasi tinklo tranzityvumo atributu. Paprastai tariant, jei viršūnė x turi jungtį su kita viršūne y, o pastaroji su dar kita viršūne z, tuomet jungties tarp viršūnių x ir z tikimybė išauga

[HCS+06]. Šis indeksas gali būti užrašomas lygybe:

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|. \tag{7}$$

- *Žakardo* (angl. Jaccard) indeksas yra normalizuota bendrų kaimynų indekso versija (žr. (8) lygybę). Normalizacijos dėka, šis indeksas atspindi tikimybę, kad iš viršūnių x ir y kaimynų atsitiktinai ir nepriklausomai pasirinkta viršūnė yra bendra x ir y kaimynė.

$$s_{xy}^{JC} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} = \frac{s_{xy}^{CN}}{|\Gamma(x) \cup \Gamma(y)|} \tag{8}$$

- *Saltono* indeksas [SM86], literatūroje dar vadinamas kosinuso panašumu [LZ11], gali būti suvokiama dar viena bendrų kaimynų indekso atmaina:

$$s_{xy}^{Salton} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \cdot |\Gamma(y)|}} = \frac{s_{xy}^{CN}}{\sqrt{|\Gamma(x)| \cdot |\Gamma(y)|}}. \tag{9}$$

- *Adamik-Adar* (angl. Adamic-Adar) indeksas [AA03] suteikia galimybę pridėti svorio bendram viršūnių poros x ir y kaimynui z atsižvelgiant į tai, kiek jungčių su kitomis viršūnėmis turi z. Dėl trupmenos vardiklio šis įvertis prideda papildomo svorio mažiau jungčių turintiems bendriems kaimynams. Adamik-Adar indeksas formaliai gali būti užrašomas lygybe:

$$s_{xy}^{AA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\lg |\Gamma(z)|}. \tag{10}$$

- *Resursų išskyrimo* (angl. resource allocation) indeksas [ZLZ09]. Šio indekso pagrindas – resursų išskyrimo ir perdavimo dinamika sudėtinguose realaus pasaulio tinkluose [LZ11]. Tarkime, turime porą tiesiogiai nesujungtų viršūnių x ir y bei šios poros bendrus kaimynus, atliekančius resursų perdavimo rolę iš x į y. Paprasčiausiu atveju darome prielaidą, kad kiekviena resursus perduodanti viršūnė šiuos resursus perduos po lygiai visiems savo kaimynams. Tokiu atveju šio indekso įvertis tarp viršūnių x ir y gali būti apibrėžtas kaip kiekis resurso, kurį viršūnė y gavo iš viršūnės x [LZ11]. Tai galima išreikšti lygybe:

$$s_{xy}^{RA} = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{|\Gamma(z)|}. \tag{11}$$

Galima pastebėti resursų išskyrimo indekso panašumą Adamik-Adar indeksą. Esant mažam kiekiui bendrų kaimynų tarp viršūnių x ir y šie indeksai iš tiesų yra labai panašūs, tačiau

bendrų kaimynų kiekiui augant rezultatas gali reikšmingai išsiskirti, kadangi resursų išskyrimo indekso atveju vardiklis auga greičiau nei logaritmo funkcija, matoma Adamik-Adar indekso išraiškoje.

1.3.1.2. Viršūnių savybių agregacija paremti metodai

Šio tipo metodai kombinuoja tam tikrus viršūnių savybių įverčius juos agreguodami, pavyzdžiui, sudedant ar sudauginant [HCS+06]. Pora tokio tipo jungties tikėtinumo įvertinimo metodų pateikiami žemiau.

- *Preferencinio prisijungimo* (angl. preferential attachment) indeksas. Preferencinio prisijungimo samprata gali būti naudojama generuoti dinamiškiems laipsninį viršūnių laipsnių pasiskirstymą turintiems tinklams, kuriuose tikimybė, kad nauja viršūnė y prisijungs prie jau esamos viršūnės x, yra proporcinga viršūnės x laipsniui (arba jo kaimynų skaičiui $|\Gamma(x)|$) [New01]. Preferencinio prisijungimo indeksas remiasi šia idėja ir gali būti išreiškiamas:

$$s_{xy}^{PA} = |\Gamma(x)| \cdot |\Gamma(y)|. \tag{12}$$

  Svarbu atkreipti dėmesį, kad vietoje sandaugos kaip agregacijos funkcija gali būti naudojama ir sudėtis [HCS+06]: $|\Gamma(x)| + |\Gamma(y)|$.

- *Klasterizacijos koeficiento agregacijos* indeksas. Klasikiniu požiūriu klasterizacijos koeficientas neorientuoto tinklo viršūnei x nusako tikimybę, kad šios viršūnės kaimynai yra taip pat kaimynai tarpusavyje [New03]. Šį dydį galima užrašyti išraiška:

$$C_x = \frac{\text{trikampių skaičius tinkle, į kuriuos įeina viršūnė } x}{\text{sujungtų trejetų skaičius, su viršūne } x \text{ viduryje}}. \tag{13}$$

  Šis tinklų atributas gali būti taikomas ir jungčių prognozei paskaičiavus jo įverčius viršūnių porai x ir y bei sudėjus arba sudauginus gautus skaičius [Puj15]:

$$s_{xy}^{C\times} = C_x \cdot C_y,\ s_{xy}^{C+} = C_x + C_y.$$

Literatūroje galima atrasti nuodugniai atliktų tyrimų, lyginančių nagrinėtų lokalių jungčių tikėtinumo indeksų prognozės kokybę per metriką AUROC (žr. (6) išraišką) [LZ11; ZLZ09]. Šių tyrimų rezultatai pateikiami lentelėje žemiau (žr. 2 lentelę).

2 lentelė. Lokalią viršūnių informaciją naudojančių jungčių tikėtinumo indeksų palyginimas per metriką AUROC (žr. (6) išraišką) realiuose tinkluose. Kiekvienas AUROC įvertis yra daugiau nei dešimties implementacijų rezultatų vidurkis su mokymo ir testavimo aibėmis, parinktomis atsitiktinai ir nepriklausomai. Didžiausios AUROC reikšmės kiekvienam tinklui paryškintos. Kiekvienas indeksas skaičiuojamas šešiuose realiuose tinkluose: (PPI) baltymų-baltymų sąveikų (angl. protein-protein interaction) tinkle, (NS) tinklų teorijos (angl. network science) mokslininkų bendraautorystės tinkle, (Grid) elektros tiekimo tinkle JAV vakaruose, (PB) JAV politinių tinklaraščių (angl. political blogs) tinkle, (INT) interneto maršrutizatorių lygmenyje tinkle, (USAir) JAV oro linijų tinkle. Parengta pagal: [LZ11], detali tinklų informacija: [ZLZ09].

Indeksas \ Tinklas	PPI	NS	Grid	PB	INT	USAir
s_{xy}^{CN}	0,889	**0,933**	**0,590**	0,925	**0,559**	0,937
s_{xy}^{JC}	0,888	**0,933**	**0,590**	0,882	**0,559**	0,901
s_{xy}^{Salton}	0,869	0,911	0,585	0,874	0,552	0,898
s_{xy}^{AA}	0,888	0,932	**0,590**	0,922	**0,559**	0,925
s_{xy}^{RA}	**0,890**	**0,933**	**0,590**	**0,931**	**0,559**	**0,955**
s_{xy}^{PA}	0,828	0,623	0,446	0,907	0,464	0,886

Kaip matyti iš lentelės, resursų išskyrimo indeksas (RA) pasirodo geriausiai prognozuojant jungtis tyrinėtuose realiuose tinkluose. Adamik-Adar (AA) ir bendrų kaimynų (CN) indeksai, tuo tarpu, užima antrą vietą. Blogiausi rezultatai matomi preferencinio prisijungimo (PA) indekso atveju. Kai kuriuose tinkluose šio metodo rezultatai yra blogesni nei atsitiktinumas (AUROC $< 0{,}5$).

1.3.2. Globalūs tinklo jungčių prognozės metodai

Naudojantis globaliais jungčių prognozės metodais, kitaip nei lokaliais, jungčių tikėtinumo indeksas skaičiuojamas žvelgiant į tinklą plačiau nei tik viršūnės ar viršūnių poros lygiu. Šis įvertis s_{xy} gali būti skaičiuojamas aibe būdų, pavyzdžiui, remiantis atstumo tarp viršūnių įvertinimu ar atsitiktiniu klaidžiojimu tinkle.

Atstumu tarp viršūnių paremti metodai iš esmės grindžiami trumpiausių arba tam tikro ilgio kelių tarp dviejų tiesiogiai nesujungtų viršūnių paieška. Šie kelių ilgiai (atstumai) yra naudojami jungties tikėtinumui arba viršūnių panašumui įvertinti. Atsitiktiniu klaidžiojimu tinkle grįsti metodai remiasi atsitiktine tam tikro ilgio kelione iš viršūnės x į viršūnę y. Žemiau pateikiama keletas dažniau sutinkamų globalių jungčių prognozės metodų.

- *Katzo* (angl. Katz) indeksas [Kat53] – vienas geriau žinomų jungčių tikėtinumo įvertinimo metodų, kuris remiasi kelių tarp viršūnių tinkle samprata. Šis metodas skaičiuoja tam tikro ilgio kelius, priskirdamas didesnius svorius trumpesniems keliams. Matematiškai tai galima

užrašyti:

$$s_{xy}^{Katz} = \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |path_{xy}^{(\ell)}|, \tag{14}$$

kur $|path_{xy}^{(\ell)}|$ yra skaičius ilgio ℓ kelių tarp viršūnių x ir y, o $\beta \in [0; 1]$ nelygumus švelninantis parametras. Katzo indeksas taip pat gali būti užrašomas ir per gretimumo matricą:

$$s_{xy}^{Katz} = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \ldots, \tag{15}$$

kur A_{xy} yra gretimumo matrica, kurios reikšmės yra 0 arba 1 priklausomai nuo to, ar viršūnes x ir y yra tiesiogiai sujungtos. $(A^{\ell})_{xy}$ yra matrica su ilgio ℓ keliais tarp viršūnių x ir y ir t.t [LZ11]. Taigi, Katzo indeksas visoms tinklo viršūnių poroms gali būti užrašomas ir matricine forma:

$$S^{Katz} = (I - \beta A)^{-1} - I, \tag{16}$$

kur I yra atitinkamo dydžio vienetinė matrica.

- *Matricų miško* (angl. matrix forest) indeksas. Šis indeksas užrašomas lygybe:

$$S^{MF} = (I + L)^{-1}, \tag{17}$$

kur I yra vienetinė matrica, o $L = D - A$ yra Laplaco matrica, išreiškiama matricų skirtumu tarp tinklo viršūnių laipsnių matricos ir jo gretimumo matricos. Remiantis šiuo metodu, jungties tikėtinumas tarp viršūnių x ir y gali būti suprantamas kaip aprėpties miškų, tokių, kad ir x, ir y patenka į ta patį aprėpties medį su viena iš šių viršūnių šaknyje, dalis nuo visų aprėpties miškų šiame tinkle [LZ11].

- *Kelionės trukmės* (angl. commute time) indeksas. Šis indeksas remiasi žingsnių, reikalingų atsitintinai tinkle klaidžiojančiam agentui nukakti iš viršūnės x į viršūnę y skaičiavimu [HCS+06]. Mažesnis žingsnių skaičius tarp viršūnių indikuoja didesnį jų panašumą ir jungties susidarymo tikėtinumą. Kadangi ši metrika nėra simetriška [Puj15], neorientuotiems tinklams ji dažnai paverčiama vidutine kelionės trukmės metrika (angl. average commute time). Jei s_{xy}^{CT} yra trukmės arba žingsnių skaičius agentui nukeliauti iš viršūnės x į viršūnę y, tada vidutinė kelionės trukmė gali būti užrašoma:

$$s_{xy}^{ACT} = s_{xy}^{CT} + s_{yx}^{CT}. \tag{18}$$

Svarbu atkreipti dėmesį į tai, kad su globalia tinklų informacija dirbantys jungčių prognozės metodai remiasi gerokai sudėtingiau algoritmiškai įvertinamais tinklų atributais, tokiais kaip keliai, atstumai ar atsitiktinis klaidžiojimas. Taip pat šie metodai yra grįsti ir ženkliai sudėtingesnėmis matematinėmis konstrukcijomis. Nepaisant to, globalūs jungčių prognozės metodai dažniausiai duoda tikslesnius rezultatus [LZ11]. Kaip to iliustraciją galima pateikti resursų išskyrimo (geriausiai pasirodžiusio lokalaus metodo (žr. 2 lentelę)) palyginimą su globaliu Katzo metodu (žr. 3 lentelę).

3 lentelė. Resursų išskyrimo (žr. (11) lygybę) ir Katzo (žr. (14) lygybę) indeksų prognozės kokybės palyginimas per metriką AUROC (žr. (6) išraišką). Iš rezultatų galima matyti Katzo indekso pranašumą visuose tinkluose. Parengta pagal: [LZ11], detali tinklų informacija: [ZLZ09].

Tinklas / Indeksas	PPI	NS	Grid	PB	INT	USAir
s_{xy}^{RA}	0,890	0,933	0,590	0,931	0,559	0,955
s_{xy}^{Katz}	**0,972**	**0,988**	**0,952**	**0,936**	**0,975**	**0,956**

1.4. Jungčių prognozė orientuotuose tinkluose

Kaip galime pastebėti iš anksčiau nagrinėtų jungčių prognozės metodų, didžioji dalis jų pirmenybę teikia neorientuotiems tinklams. Ši situacija ir faktas, kad orientuoti tinklai yra plačiai sutinkami realiame pasaulyje, iš esmės motyvuoja atlikti nuodugnesnius jungčių prognozės tyrimus juose. Kaip orientuotų tinklų pavyzdžius galima paminėti mitybos grandines, citavimo tinklus ar didelę dalį internetinių socialinių tinklų, kuriuose vyrauja, pzvydzžiui, sekėjo (angl. follower) ir sekamojo idėja (pvz. *Twitter*, *Instagram*[5]).

Jungčių prognozė orientuotuose tinkluose yra ganėtinai sudėtingas uždavinys (trys viršūnės gali sudaryti trylika skirtingų orientuotų tinklų!). Pagrindinė šio sudėtingumo priežastis čia yra briaunų kryptingumas. Dėl šios tinklo savybės tokie metodai, kaip, tarkime, anksčiau nagrinėtas bendrų kaimynų indeksas be papildomų modifikacijų orientuoto tinklo viršūnių porai paprasčiausiai negali būti įvertintas. Taip yra todėl, kadangi šioms viršūnėms nėra apibrėžta kaimyno samprata (kuria kryptimi yra kaimynai?). Jungčių prognozės metodai šiems tinklams turi atsižvelgti į tinklo motyvus ir šiuose motyvuose glūdinčią tinklo informaciją. Kitu atveju, net ir nuspėjus jungtį tarp viršūnių poros nebūtų galima nustatyti šios jungties krypties [LZ11]. Toliau

[5] `https://www.instagram.com/`

šiame darbe dėmesys sukoncentruojamas būtent į orientuotus tinklus ir jungčių prognozę juose. Esminė idėja, kuria toliau remiamasi šiame darbe yra orientuoto tinklo klasterizacijos samprata.

* * *

Tinklo jungčių prognozės skyriuje formaliai apibrėžėme ir išnagrinėjome tinklo jungčių prognozės uždavinį bei jo tipus. Remdamiesi literatūra, taip pat apibrėžėme karkasą šio uždavinio sprendimo metodų vertinimui bei pateikėme aibę tam taikytinų metrikų. Vėliau išanalizavome keletą dažniau sutinkamų jungčių prognozės metodų, kuriuos sugrupavome į dvi grupes: lokalius ir globalius pagal tai, su kokio lygio informacija tinkle jie dirba (viršūnių, potinklio ar viso tinklo lygio informacija). Be to, pateikėme ir metodų prognozės kokybės eksperimentinius įverčius realiuose tinkluose bei juos lyginome tarpusavyje su tikslu įvertinti nagrinėtų metodų prognozės kokybę. Svarbu turėti omenyje, kad be mūsų pristatytų toploginių tinklo jungčių prognozės būdų egzistuoja dar gausybė įvairiausių kitokių metodų, apimančių tikimybinius, hierarchinius, stochastnius modelius [LZ11; Puj15]. Šie metodai yra nemažiau svarbūs sprendžiant jungčių prognozės uždavinius, tačiau yra mažiau reikšmingi šiame darbe nagrinėjamiems klausimams. Paskutiniame poskyryje išskyrėme ir esminius iššūkius jungčių prognozei orientuotuose tinkluose bei padėjome pagrindą su tuo susijusiai tolesnei darbo daliai.

2. Klasterizacijos samprata tinkluose

Daugelyje realaus pasaulio tinklų stebima viršūnių tendencija grupuotis į tankias lokalias bendruomenes (angl. clusters) globaliu požiūriu retame tinkle (esamos jungtys tinkle sudaro tik nedidelę dalį visų galimų jungčių). Ši savybė įprastai įvertinama klasterizacijos koeficientu, nusakančiu, kiek tikėtina, kad viršūnės kaimynės tarpusavyje irgi yra kaimynės [BL16].

2.1. Klasterizacija neorientuotuose tinkluose

Neorientuotuose tinkluose klasterizacijos samprata remiasi prielaida, kad, jeigu viršūnė x turi jungtį (yra kaimynė) su viršūne y, o pastaroji – su viršūne z, tuomet išauga tikimybė, kad viršūnė x taip pat yra sujungta su viršūne z. Žiūrint iš socialinių tinklų perspektyvos, mažiau formaliai tai galima nusakyti šitaip: tavo draugų draugai, tikėtina, tau taip pat yra draugai [New03]. Tinklų topologijos terminologija šį reiškinį galima įvardinti kaip tranzityvumą, atspindintį trikampių kiekį tinkle. Trikampis čia suvokiamas kaip trijų viršūnių, sujungtų briaunomis rinkinys (K_3 grafas). Ši savybė neorientuotuose tinkluose dažnai kiekybiškai vertinama klasterizacijos koeficientu, kuris gali būti skaičiuojamas globaliai (visam tinklui) ir lokaliai (vienai viršūnei). Globalųjį atvejį galima užrašyti kaip santykį [New03]:

$$C = \frac{3 \cdot \text{trikampių skaičius tinkle}}{\text{sujungtų trejetų skaičius}}, \tag{19}$$

kur „sujungtas trejetas“ reiškia viršūnę, tiesiogiai sujungtą su pora jai kaimyninių viršūnių, čia tvarka nėra svarbi. Kitaip tariant, šis santykis parodo, kokia dalis sujungtų tinklo viršūnių trejetų sudaro pilnus trikampius (K_3 pografius).

Kadangi kiekvieną trikampį neorientuotame tinkle sudaro trys skirtingi viršūnių trejetai, santykio (19) skaitiklis dauginamas iš trijų. Tokiu būdu užtikrinama, kad globalusis klasterizacijos koeficientas patenka į intervalą $[0; 1]$. Čia svarbu pastebėti tai, jog dydį C galima interpretuoti ir kaip tikimybę, kad bet kokios dvi tinklo viršūnės, turinčios bendrą kaimyninę viršūnę, taip pat yra tarpusavyje kaimynės (yra susietos briauna) [New03].

Globalųjį klasterizacijos koeficientą neorientuotiems tinklams galima išreikšti ir per kelių tinkluose sampratą:

$$C = \frac{6 \cdot \text{trikampių skaičius tinkle}}{\text{keliai, kurių ilgis yra du skaičius}}. \tag{20}$$

Šiuo atveju skaitiklyje matomas šešetas reikalingas todėl, kad kalbant apie kelius tinkle yra svarbi tvarka ir kiekvieną sujungtą trejetą atitinka du skirtingi keliai.

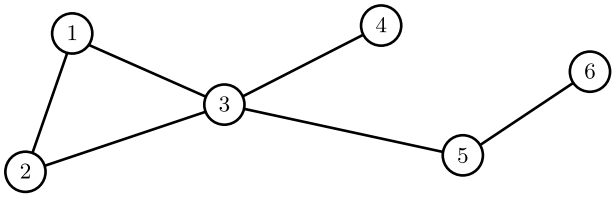
Lokalusis klasterizacijos koeficientas C_x neorientuoto tinklo viršūnei x yra gali būti išreiškiamas santykiu [WS98]:

$$C_x = \frac{\text{trikampių skaičius tinkle, į kuriuos įeina viršūnė } x}{\text{sujungtų trejetų skaičius, su viršūne } x \text{ viduryje}}. \tag{21}$$

Reikėtų atkreipti dėmesį, kad, jeigu viršūnės x laipsnis (kaimynų skaičius) yra 0 arba 1 (santykio C_x vardiklis ir skaitiklis tampa nuliais), lokalusis klasterizacijos koeficientas sutartinai laikomas nuliu [New03]. Naudojantis lokaliu koeficiento įverčiu, nesunkiai paskaičiuojamas ir vidutinis jo įvertis, kuris visam tinklui su n viršūnių gali būti užrašomas suma:

$$\overline{C} = \frac{1}{n}\sum_{x=1}^{n} C_x. \tag{22}$$

Nesunku pastebėti, kad vidutinis klasterizacijos koeficientas skaičiuoja trikampių ir trejetų santykių vidurkį, kai globalusis – šių vidurkių santykį. Dėl šios priežasties abiejų koeficientų įverčiai kartais gali šiek tiek skirtis [New03]. Tiek vienas, tiek kitas klasterizacijos koeficientas yra plačiai sutinkamas literatūroje. Globalusis atvejis dažnai patogesnis matematinei tinklų analizei dėl tikimybinės prasmės, vidutinis – patogus tuo, kad yra nesunkiai algoritmiškai įvertinamas.



4 pav. Klasterizacijos koeficiento skaičiavimo neorientuotiems tinklams pavyzdys. Vaizduojamas tinklas turi vieną trikampį (1,2,3) ir devynis sujungtus viršūnių trejetus: $(1,2,3), (2,3,1), (2,3,4), (2,3,5), (2,1,3), (3,4,5), (1,3,4), (1,3,5), (3,5,6)$. Globalusis klasterizacijos koeficientas šiam tinklui $C = \frac{3 \cdot 1}{9} = \frac{1}{3}$. Lokalūs kiekvienai viršūnei klasterizacijos koeficiento įverčiai yra: $C_1 = C_2 = 1, C_3 = \frac{1}{6}, C_4 = C_5 = C_6 = 0$. Atitinkamai, vidutinis viso tinklo klasterizacijos koeficiento įvertis $\overline{C} = \frac{13}{36}$. Šiuo atveju $\overline{C}$ yra neženkliai didesnis už C.

2.2. Klasterizacija orientuotuose tinkluose

Orientuotuose tinkluose klasterizacijos samprata tampa gerokai sudėtingesnė dėl tinklo jungčių kryptingumo. Pagrindinis iššūkis čia – prasmingas šio kryptingumo panaudojimas nagrinėjant klasterizaciją [FV13]. Be to, vis dar nėra nusistovėjusios ar pakankamai plačiai pripažintos klasterizacijos orientuotuose tinkluose sampratos kaip neorientuotų tinklų atveju [FV13].

Šiandien jau yra išvystytas ne vienas metodas, savaip apibrėžiantis klasterizaciją orientuotuose tinkluose. Dalis jų remiasi tinklų transformacijomis, dalis tikimybiniais ar informacijos teorijos modeliais, dalis yra grįsti tam tikrų tinklo motyvų paieška ir analize [FV13].

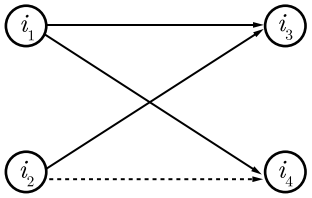
2.3. Diklikos klasterizacijos koeficientas

Šiame darbe nagrinėjama M. Bloznelio ir L. Leskelos (suom. Leskelä) pristatyta nauja orientuoto tinklo klasterizacijos samprata [BL16]. Šios sampratos esminė motyvacija, pasak autorių, yra orientuoti socialiniai tinklai, kuriuose orientuota jungtis tarp tinklo viršūnių (esybių) x ir y iliustruoja faktą, kad viršūnė x seka[6] viršūnę y ($x \to y$). Šios sampratos idėją tokiuose tinkluose neformaliai galima iliustruoti teiginiu: *viršūnės sekėjai, tikėtina, turi dar ir kitų bendrai sekamų viršūnių* [BL16].

Tokio tipo klasterizacijos koeficientas orientuotuose tinkluose gali būti suvokiamas kaip sąlyginė tikimybė, kad dvi viršūnės, turinčios orientuotą briauną į bendrą viršūnę, turės bendrų jungčių į dar bent vieną kitą tinklo viršūnę. Tinkle, kurio viršūnių laipsniai yra pasiskirstę pagal skirstinį $\mathbf{P}$, ši tikimybė gali būti būti užrašoma [BL16]:

$$\mathbf{P}(i_2 \to i_4 | i_1 \to i_3, i_1 \to i_4, i_2 \to i_3), \tag{23}$$

kur viršūnė i_3 yra sekamoji viršūnė. Tai patogu įsivaizduoti kaip grafą (žr. 5 pav.), dar žinomą kaip dikliką (angl. diclique). Toks tinklo motyvas, kaip pastebima, dominuoja daugybėje realaus pasaulio tinklų, tokių kaip, pavyzdžiui, citavimo ar internetiniai socialiniai tinklai [ZLW+13].



5 pav. Tinklo motyvas, iliustruojantis tikimybės (23) išraiškoje idėją. Egzistuojant jungčiai $i_2 \to i_4$ šis tinklo motyvas tampa diklika. Parengta pagal: [BL16].

Remdamiesi aukščiau nagrinėjama orientuoto tinklo klasterizacijos samprata, autoriai apibrėžia globalų *diklikos klasterizacijos koeficientą* baigtiniam orientuotam tinklui D su gretimumo matrica D_{ij}. Šį koeficientą galima išreikšti santykiu:

$$C(D) = \frac{\sum_{(i_1,i_2,i_3,i_4)} D_{i_1,i_3} D_{i_1,i_4} D_{i_2,i_3} D_{i_2,i_4}}{\sum_{(i_1,i_2,i_3,i_4)} D_{i_1,i_3} D_{i_1,i_4} D_{i_2,i_3}}, \tag{24}$$

kur skaitiklio ir vardiklio sumos yra skaičiuojamos visiems skirtingiems tinklo viršūnių ketvertams (tvarka čia svarbi). Ši išraiška gali būti suprantama, kaip labiau praktinis sąlyginės tikimybės (23) išraiškoje atitikmuo, nusakantis, kokią dalį nuo visų skirtingų (tvarka svarbi) tinklo D viršūnių

[6] Sekimas (angl. following) socialiniuose tinkluose, tokiuose kaip *Twitter*, atspindi vienpusį ryšį tarp dviejų tinklo viršūnių (naudotojų), iš kurių vienas prenumeruoja (stebi, domisi) kito skelbiama informacija, naujienomis ir t.t.

ketvertų (diklikų arba ketvertų, kuriems iki diklikos trūksta jungties $i_2 \to i_4$) sudaro pilnos diklikos. Šis dydis įvertina kiek kitokią sąlyginę tikimybę [BL16]:

$$\mathbf{P}_D(I_2 \to I_4 | I_1 \to I_3, I_1 \to I_4, I_2 \to I_3), \tag{25}$$

kur $\mathbf{P}_D$ yra skirstinys atsitiktinių ketvertų (I_1, I_2, I_3, I_4), pasirenkamų tolygiai atsitiktinai iš visų skirtingų viršūnių ketvertų tinkle D (tvarka yra svarbi).

Lokaliai tinklo D viršūnei i nagrinėjamas klasterizacijos koeficientas gali būti išreiškiamas:

$$C(D, i) = \frac{\sum_{(i_1,i_2,i_4)} D_{i_1,i} D_{i_1,i_4} D_{i_2,i} D_{i_2,i_4}}{\sum_{(i_1,i_2,i_4)} D_{i_1,i} D_{i_1,i_4} D_{i_2,i}}, \tag{26}$$

kur vardiklio ir skaitiklio sumos yra įvertinamos visiems skirtingiems tinklo trejetams be viršūnės i (tvarka svarbi). Taip pat reikėtų atkreipti dėmesį, kad sąlyginė tikimybė (25) čia tampa [BL16]:

$$\mathbf{P}_D(I_2 \to I_4 | I_1 \to I_3, I_1 \to I_4, I_2 \to I_3, I_3 = i). \tag{27}$$

2.4. Diklikos klasterizacijos koeficiento praplėtimai

Aukščiau pristatytas diklikos klasterizacijos koeficientas $C(D)$ orientuotam tinklui D praplečiamas į tris papildomus atvejus [Dav17; Vai17] (labiau orientuojantis į jungčių prognozės sritį):

1. Koeficientas $C^{IN}(D,k)$ atsižvelgia tik į tuos viršūnių ketvertus, kuriuose sekamosios viršūnės i_3 įėjimo laipsnis (angl. indegree) yra lygus k. Šį koeficiento papildymą, remiantis aukščiau apibrėžta sąlygine tikimybe (žr. (25) išraišką), galima užrašyti kaip:

$$\mathbf{P}_D(I_2 \to I_4 | I_1 \to I_3, I_1 \to I_4, I_2 \to I_3, d_{in}(I_3) = k). \tag{28}$$

   Koeficiento priklausomybę nuo sekamosios viršūnės įėjimo laipsnio patogu užrašyti išraiška:

$$C^{IN}(D,k) = \frac{N_1(D,k)}{N_1(D,k) + N_2(D,k)}, \tag{29}$$

   kur $N_1(D,k)$ yra grafo D pografių, sudarančių dikliką, skaičius, o $N_2(D,k)$ – grafo D pografių, kuriems iki diklikos trūksta tik jungties $i_2 \to i_4$, skaičius. Abiem atvejais turi būti tenkinama sąlyga $d_{in}(i_3) = k$.

2. Koeficientas $C^W(D,w)$ įveda jungties liudininko sampratą. Liudininku čia laikome tokią viršūnę, kuri turi orientuotą briauną į mus dominančios jungties galutinę viršūnę bei da-

linasi bendrai sekamomis viršūnėmis su mus dominančios jungties pradine viršūne. Kaip pavyzdį tam iliustruoti galima panaudoti grafą 5 paveikslėlyje. Mus domina ryšys $i_2 \to i_4$, apie kurį informacijos mes neturime. Yra žinoma, kad viršūnė i_2 turi bent vieną bendrai sekamą viršūnę su i_1 ($i_2 \to i_3, i_1 \to i_3$), kuri turi briauną į mus dominančios jungties galutinę viršūnę i_4 ($i_1 \to i_4$). Dėl to viršūnė i_1 yra laikoma jungties $i_2 \to i_4$ liudininke. Be to, koeficientas $C^W(D,w)$ atsižvelgia tik į tuos atvejus, kada dominanti jungtis turi lygiai w liudininkų. Šio koeficiento priklausomybę nuo liudininkų skaičiaus w išreiškiame taip:

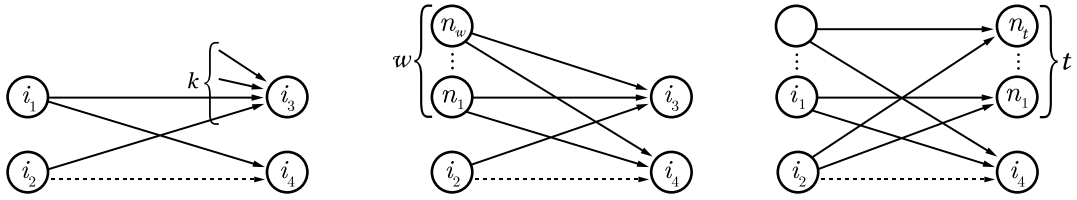
$$C^W(D,w) = \frac{W_1(D,w)}{W_1(D,w) + W_2(D,w)}, \tag{30}$$

kur $W_1(D,w)$ yra grafo D liudininkų pografių, turinčių jungtį $i_2 \to i_4$, skaičius, o $W_2(D,w)$ – skaičius pografių, neturinčių jungties $i_2 \to i_4$. Abiem atvejais turi būti tenkinama sąlyga $|W(i_2,i_4)| = w$, kur $W(i_2,i_4)$ yra jungties $i_2 \to i_4$ liudininkų aibė.

3. Koeficientas $C^T(D,t)$ yra šiek tiek panašus koeficientą $C^W(D,w)$. Čia naudojamasi bendrai sekamų viršūnių (taikinių) skaičiumi t. Bendrai sekama viršūne šiuo atveju yra laikoma ta viršūnė, kuri yra bendrai sekama mus dominančios jungties pradinės viršūnės ir jos galutinės viršūnės sekėjų. Pasinaudoję grafu 5 paveikslėlyje matome, jog viršūnė i_3 yra viršūnių i_2 ir i_1 bendrai sekama viršūnė, o i_1 seka mus dominančios jungties $i_2 \to i_4$ galutinę viršūnę i_4. Šio koeficiento priklausomybę nuo i_2 ir i_1 bendrų taikinių skaičiaus t išreiškiame:

$$C^T(D,t) = \frac{T_1(D,t)}{T_1(D,t) + T_2(D,t)}, \tag{31}$$

kur $T_1(D,t)$ yra grafo D taikinių pografių, turinčių jungtį $i_2 \to i_4$, skaičius, o $T_2(D,t)$ – skaičius pografių, neturinčių jungties $i_2 \to i_4$. Abiem atvejais turi būti tenkinama sąlyga $|T(i_1,i_2)| = t$, kur $T(i_1,i_2)$ yra i_1 ir i_2 bendrai sekamų viršūnių aibė.



(a) $C^{IN}(D,k)$ (b) $C^W(D,w)$ (c) $C^T(D,t)$

6 pav. Diklikos klasterizacijos koeficientų praplėtimų idėjas iliustruojantys tinklo motyvai. Svarbu atkreipti dėmesį, jog šiuose motyvuose atvaizduotos tik tos jungtys, kurios daro įtaką koeficientų įverčiams. Kitos jungtys, tokios kaip $i_1 \to i_2$, egzistuoti gali, bet jos mūsų nedomina.

Aukščiau pristatytų diklikos klasterizacijos koeficientų praplėtimai įveda kintamuosius k, w ir t tokiu būdu atitinkamai paversdami šiuos koeficientus priklausomais kintamaisiais – funkcijomis. Šios funkcijos reprezentuoja klasterizacijos koeficientų priklausomybę nuo minėtųjų kintamųjų, o patys kintamieji, savo ruožtu, atspindi tam tikrus tinklo topologijos atributus (viršūnių įėjimo laipsnį, liudininkų bei bendrai sekamų viršūnių skaičius), galinčius turėti reikšmingą įtaką šio tinklo jungčių susidarymui. Visi trys koeficientai vėliau darbe naudojami jungčių prognozei, o minėtosios priklausomybės mums suteikia šiek tiek daugiau konteksto bei padeda įžvelgti tam tikras viršūnių formavimosi tendencijas.

3. Prognozė, paremta diklikos klasterizacijos koeficientu

Kaip jau pastebėjome anksčiau, diklikos klasterizacijos koeficientas yra apibrėžiamas kaip sąlyginė briaunos egzistavimo tikimybė (žr. (23) išraišką). Tai natūraliai sufleruoja galimą šio koeficiento tinkamumą jungtims orientuotuose tinkluose prognozuoti. To motyvuojami šiame darbe ir siekiame įvertinti diklikos klasterizacijos koeficiento tinkamumą jungčių prognozės uždaviniams spręsti.

3.1. Jungčių tikėtinumo indeksai

Naudodamiesi praeitame skyriuje suformuluotais diklikos klasterizacijos koeficiento praplėtimais, atitinkamai apibrėžiame tris orientuoto tinklo D jungties tikėtinumo indeksus šio tinklo viršinių porai (x,y).

1. *Įėjimo laipsnių agregacijos* indeksas. Diklikos klasterizacijos koeficiento praplėtimu $C^{IN}(D,k)$ paremtas jungties tikėtinumo indeksas. Šis indeksas gali būti užrašomas lygybe:

$$s_{xy}^{CIN} = \sum_{z \in T(x,y)} C^{IN}(D, d_{in}(z)), \tag{32}$$

   kur $T(x,y)$ yra tinklo D viršūnės x bendrai su viršūnės y sekėjais sekamų viršūnių aibė. Koeficientų suma čia pasirinkta todėl, kad aibė $T(x,y)$ gali turėti daug elementų. Čia galėtų būti naudojama ir sandauga, tačiau sudėtis įprastai indikuoja tikimybinį nepriklausomumą.

2. *Liudininkų* indeksas. Diklikos klasterizacijos koeficiento praplėtimu $C^{W}(D,w)$ paremtas jungties tikėtinumo indeksas. Šis indeksas gali būti užrašomas lygybe:

$$s_{xy}^{CW} = C^{W}(D, |W(x,y)|), \tag{33}$$

   kur $W(x,y)$ tinklo D briaunos (x,y) egzistavimo liudininkų aibė.

3. *Bendrų interesų* indeksas. Diklikos klasterizacijos koeficiento praplėtimu $C^{T}(D,t)$ paremtas jungties tikėtinumo indeksas. Šis indeksas gali būti užrašomas lygybe:

$$s_{xy}^{CT} = C^{T}(D, |T(x,y)|), \tag{34}$$

   kur $T(x,y)$, kaip ir pirmuoju atveju, yra tinklo D viršūnės x bendrai su viršūnės y sekėjais sekamų viršūnių (taikinių) aibė.

4. Tyrimo duomenų surinkimas

Jungčių prognozės sėkmė dažnai stipriai priklauso ir nuo tiriamo tinklo, kuriame ji yra atliekama. Tai nulemia paties tinklo ir prognozės metodo specifika. Dėl šios priežasties jungčių prognozės įrankį svarbu išmėginti su kuo įvairesniais tinklais. Tokiu būdu siekiama kiek įmanoma mažiau šališkai įvertinti tiriamo metodo prognozavimo kokybę.

Šiame darbe anksčiau apibrėžti jungčių tikėtinumo indeksai yra tiriami su kelių tipų realiais orientuotais tinklais. Nors dėl tiriamų metodų teorinio pagrindo prigimties darbe orientuojamasi į socialinius tinklus, jungčių prognozės metodai išmėginami su poros skirtingų sričių tinklais (socialiniais ir informaciniais). Visi tiriami tinklai taip pat yra ir skirtingo dydžio bei tankumo (viršūnių ir briaunų prasme atitinkamai). Tai suteikia galimybę įvertinti tiriamų prognozės metodų kokybę tais atvejais, kada turima mažai ir daug informacijos apie tinklo topologiją. Dalis nagrinėjamų tinklų yra žinomi ir nemažai tinklų teorijoje analizuoti tinklai, dalis – specialiai šiam tyrimui iš surinktų duomenų sumodeliuoti tinklai.

Šiame darbe nagrinėjamus tinklus galima suklasifikuoti į dvi kategorijas, atspindinčias du tinklo jungčių prognozės skyriuje pristatytus jungčių prognozės uždavinio tipus. Pirmoji darbe tiriamų tinklų kategorija yra nuo laiko nepriklausomi statiniai tinklai (struktūrinė jungčių prognozė). Čia svarbu atkreipti dėmesį, jog šie statiniai tinklai iš prigimties yra nebūtinai nekintantys laike. Omenyje turima tai, kad apie šiuos tinklus tyrimo metu nėra žinoma jokia laikinė informacija, leidžianti juos nagrinėti jų evoliucijos kontekste. Antroji tiriamų tinklų kategorija – su laiku evoliucionuojantys tinklai (laikinė jungčių prognozė). Ši kategorija apima visus specialiai šiam darbui sumodeliuotus tinklus. Svarbu paminėti ir tai, kad visi tiriami tinklai yra orientuoti paprastieji grafai be pasikartojančių briaunų ir kilpų (ši informacija mūsų nedomina).

4.1. Tinklai be laiko požymio

Žemiau pateikiamas sąrašas darbe tiriamų orientuotų tinklų be laiko požymio su trumpais jų aprašymais. Visi tinklai iš tiesų yra dinamiški ir kintantys laike, tačiau tyrimo metu ši informacija mums yra nežinoma t.y. duomenų rinkiniai nesuteikia galimybės gauti tinklo jungčių susiformavimo laiko požymio.

- **WIK** (*Wikipedia Vote*[7]) [LHK10]. *Wikipedia*[8] yra nemokama elektroninė enciklopedija, kuriama savanorių iš viso pasaulio. Dalis šios enciklopedijos savanorių yra administratoriai, kurie yra išrenkami balsavimo arba diskusijos bendruomenėje būdu. Tinklas *Wikipedia Vote*

[7] `https://snap.stanford.edu/data/wiki-Vote.html`
[8] `http://www.wikipedia.org`

yra sukonstruotas iš *Wikipedia* administratorių rinkimų duomenų nuo pat enciklopedijos atsiradimo dienos iki 2008 metų sausio trečiosios. Gautas tinklas apima 2794 rinkimus su 7115 savanorių. Viršūnė šiame tinkle vaizduoja savanorį, dalyvavusį administratoriaus rinkimuose, o briauna iš viršūnės i į viršūnę j liudija faktą, kad savanoris i balsavo už savanorį j.

- **SCC** (*Statisticians' Coauthorship*[9]) [PJ14]. Originali duomenų aibė, pagal kurią sukonstruotas šis tinklas, yra dvidalis grafas, susidedantis iš 3607 statistikų (mokslininkų) ir 3248 mokslinių straipsnių. Taip pat yra žinomi kiekvieno straipsnio autoriai bei kiekvieną straipsnį cituojantys kiti straipsniai. Orientuotas tinklas šiuo atveju konstruojamas viršūnėmis pasirenkant autorius. Kryptinis ryšys iš autoriaus i į autorių j sukuriamas tuo atveju, jei autoriaus i parašytas straipsnis cituoja autoriaus j parašytą straipsnį. Gautas tinklas nėra paprastasis grafas, kadangi jame yra galimos kilpos (authorius i gali cituoti savo parašytą straipsnį). Tiriant šį tinklą kilpos yra ignoruojamos.

4.2. Tinklai su laiko požymiu

Žemiau pateikiamas darbe tiriamų orientuotų tinklų su laiko požymiu sąrašas kartu su šių tinklų aprašymu. Kiekvienas iš žemiau pateikiamų tinklų yra sumodeliuotas specialiai šiame darbe atliekamam tyrimui. Be to, kiekviena tinklo briauna turi požymį, indikuojantį šios briaunos susidarymo laiką.

Visi žemiau pateikiami tinklai yra sukonstruoti iš *StackExchange*[10] duomenų[11]. Šie duomenys nuo 2009 metų yra periodiškai atnaujinami kartą per kelis mėnesius. *StackExchange* yra atvira klausimų ir atsakymų platforma įvairiausiomis temomis nuo inžinerijos ar matematikos, iki lingvistikos, dizaino ar kovos menų. Patogiam duomenų valdymui darbo metu sukurtas programinis įrankis, skirtas iš šių duomenų konstruoti tinklus. Minėtojo įrankio pagalba iš pradinių duomenų, paimtų 2018 metų lapkričio 12 dieną, sukonstruoti šie tinklai:

- **SPA** (*StackExchange Physics Answers*). Tai yra klausimų ir atsakymų tinklas fizikos[12] tematika. Duomenų aibė, iš kurios sukonstruotas šis tinklas, yra trejetas dvidalių tinklų:

    1. Pirmasis tinklas apima bent vieną klausimą uždavusius portalo naudotojus, klausimus ir orientuotas jungtis su laiko žyme iš naudotojų į jų parašytus klausimus.

[9] `https://arxiv.org/abs/1410.2840`
[10] `https://stackexchange.com`
[11] `https://archive.org/download/stackexchange`
[12] `https://physics.stackexchange.com`

2. Antrasis tinklas susideda iš bent vieną atsakymą pateikusių portalo naudotojų, atsakymų ir orientuotų jungčių iš naudotojų į jų pateiktus atsakymus.

3. Trečiasis tinklas apima atsakymus, klausimus ir orientuotas jungtis iš kiekvieno atsakymo į klausimą, kurį jis atsako.

Iš šio trejeto sukonstruotame naudotojų tinkle viršūnės vaizduoja portalo naudotojus, kurie vieni kitiems užduoda klausimus ir siūlo atsakymus į klausimus. Kryptinis ryšys iš naudotojo i į naudotoją j šiame tinkle atsiranda tada, kada naudotojas i parašo atsakymą į naudotojo j klausimą. Šį tinklą sudaro 48445 viršūnės (naudotojai). Nors originaliame duomenų rinkinyje yra daugiau nei milijonas naudotojų, reikšminga jų dalis pašalinama, kadangi jie yra izoliuoti (nesąveikauja su kitais naudotojais). Dėl tyrimo specifikos pasikartojančios briaunos (naudotojas gali atsakyti į tą patį klausimą kelis kartus) ir kilpos (naudotojas gali atsakyti savo klausimus) yra pašalintos.

4 lentelė. Dvidalių *StackExchange* klausimų-atsakymų tinklų fizikos tematika duomenys. Čia $|V|$ ir $|E|$ atitinkamai žymi tinklo viršūnių ir briaunų skaičius. A ir B žymi dvidalio tinklo dalių viršūnių aibes ($V = A \cup B$). Dydžiai $\overline{d}_A$ ir $\overline{d}_B$ vaizduoja vidutinius šių dalių viršūnės laipsnius, o dydis $\overline{d}$ žymi bendrą dvidalio tinklo vidutinį viršūnės laipsnį.

Tinklas	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert A\rvert$	$\lvert B\rvert$	$\overline{d}$	$\overline{d}_A$	$\overline{d}_B$
Naudotojai (A) → Klausimai (B)	163297	121311	41986	121311	1,49	2,89	1,00
Naudotojai (A) → Atsakymai (B)	197082	176612	20470	176612	1,79	8,63	1,00
Atsakymai (A) → Klausimai (B)	283668	181242	181242	102426	1,28	1,00	1,77

- **SPC** (*StackExchange Physics Comments*). Tai yra klausimų ir atsakymų komentarų tinklas fizikos tematika. Klausimai ir atsakymai toliau bendrai vadinami įrašais. Duomenų aibė, iš kurios sukonstruotas šis tinklas, yra trejetas dvidalių tinklų:

  1. Pirmasis tinklas apima bent vieną įrašą paskelbusius portalo naudotojus, įrašus (klausimus ir atsakymus bendrai) ir orientuotas jungtis su laiko žyme iš naudotojų į jų sukurtus įrašus (klausimus ir atsakymus).

  2. Antrasis tinklas susideda iš bent vieną įrašą pakomentavusių portalo naudotojų, komentarų ir orientuotų jungčių iš naudotojų į jų parašytus komentarus.

  3. Trečiasis tinklas apima komentarus, įrašus (klausimus ir atsakymus bendrai) ir orientuotas jungtis iš kiekvieno komentaro į įrašą (klausimą arba atsakymą), kurį jis komentuoja.

Iš šio trejeto sukonstruotame naudotojų tinkle viršūnės vaizduoja portalo naudotojus, kurie vieni kitiems užduoda klausimus, siūlo atsakymus ir komentuoja vieni kitų įrašus (klausimus bei atsakymus). Kryptinis ryšys iš naudotojo i į naudotoją j šiame tinkle atsiranda tada, kada naudotojas i pakomentuoja naudotojo j įrašą (klausimą arba atsakymą). Šį tinklą sudaro 45541 viršūnė (naudotojai). Nors originaliame duomenų rinkinyje yra daugiau nei milijonas naudotojų, reikšminga jų dalis pašalinama, kadangi jie yra izoliuoti (nesąveikauja su kitais naudotojais). Dėl tyrimo specifikos pasikartojančios briaunos (naudotojas gali komentuoti tą patį įrašą kelis kartus) ir kilpos (naudotojas gali komentuoti savo įrašus) yra pašalintos.

5 lentelė. Dvidalių *StackExchange* komentarų tinklų fizikos tematika duomenys. Čia $|V|$ ir $|E|$ atitinkamai žymi tinklo viršūnių ir briaunų skaičius. A ir B žymi dvidalio tinklo dalių viršūnių aibes ($V = A \cup B$). Dydžiai $\overline{d}_A$ ir $\overline{d}_B$ vaizduoja vidutinius šių dalių viršūnės laipsnius, o dydis $\overline{d}$ žymi bendrą dvidalio tinklo vidutinį viršūnės laipsnį.

Tinklas	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert A\rvert$	$\lvert B\rvert$	$\overline{d}$	$\overline{d}_A$	$\overline{d}_B$
Naudotojai (A) → Klausimai ir atsakymai (B)	351977	297923	54054	297923	1,69	5,51	1,00
Naudotojai (A) → Komentarai (B)	624553	591555	32998	591555	1,89	17,93	1,00
Komentarai (A) → Klausimai ir atsakymai (B)	787786	611568	611568	176218	1,55	1,00	3,47

- **SSA** (*StackExchange Statistics Answers*). Tai yra klausimų ir atsakymų tinklas statistikos[13] tematika. Šios tinklo viršūnės ir briaunos reprezentuoja tas pačias esybes ir reiškinius kaip ir **SPA** tinkle. Tinklas taip pat ir konstravimo prasme yra analogiškas tinklui **SPA** tik atspindi kiek kitokio pobūdžio ir kiek didesnę duomenų aibę (49869 viršūnės).

6 lentelė. Dvidalių *StackExchange* klausimų-atsakymų tinklų statistikos tematika duomenys. Čia $|V|$ ir $|E|$ atitinkamai žymi tinklo viršūnių ir briaunų skaičius. A ir B žymi dvidalio tinklo dalių viršūnių aibes ($V = A \cup B$). Dydžiai $\overline{d}_A$ ir $\overline{d}_B$ vaizduoja vidutinius šių dalių viršūnės laipsnius, o dydis $\overline{d}$ žymi bendrą dvidalio tinklo vidutinį viršūnės laipsnį.

Tinklas	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert A\rvert$	$\lvert B\rvert$	$\overline{d}$	$\overline{d}_A$	$\overline{d}_B$
Naudotojai (A) → Klausimai (B)	185674	128208	57466	128208	1,38	2,23	1,00
Naudotojai (A) → Atsakymai (B)	142718	124832	17886	124832	1,75	6,98	1,00
Atsakymai (A) → Klausimai (B)	209034	124832	124832	84202	1,19	1,00	1,48

- **SSC** (*StackExchange Statistics Comments*). Tai yra klausimų ir atsakymų komentarų tinklas statistikos tematika. Šios tinklo viršūnės ir briaunos reprezentuoja tas pačias esybes ir reiškinius kaip ir **SPC** tinkle. Tinklas taip pat ir konstravimo prasme yra analogiškas tinklui **SPC** tik atspindi kiek kitokio pobūdžio ir kiek didesnę duomenų aibę (53462 viršūnės).

[13] `https://stats.stackexchange.com`

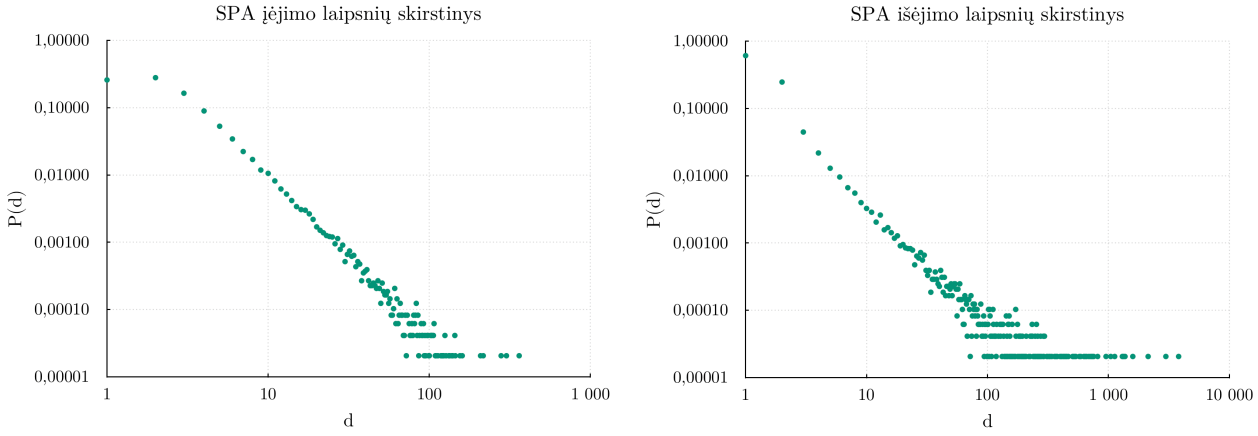
7 lentelė. Dvidalių *StackExchange* komentarų tinklų statistikos tematika duomenys. Čia $|V|$ ir $|E|$ atitinkamai žymi tinklo viršūnių ir briaunų skaičius. A ir B žymi dvidalio tinklo dalių viršūnių aibes ($V = A \cup B$). Dydžiai $\overline{d}_A$ ir $\overline{d}_B$ vaizduoja vidutinius šių dalių viršūnės laipsnius, o dydis $\overline{d}$ žymi bendrą dvidalio tinklo vidutinį viršūnės laipsnį.

Tinklas	$\|V\|$	$\|E\|$	$\|A\|$	$\|B\|$	$\overline{d}$	$\overline{d}_A$	$\overline{d}_B$
Naudotojai (A) → Klausimai ir atsakymai (B)	316979	249946	67033	249946	1,58	3,73	1,00
Naudotojai (A) → Komentarai (B)	512983	472505	40478	472505	1,84	11,67	1,00
Komentarai (A) → Klausimai ir atsakymai (B)	623614	479643	479643	479641	1,54	1,00	3,33

8 lentelė. Jungčių prognozės tyrime naudojamų tinklų duomenys. Čia $|V|$ žymi tinklo viršūnių skaičių, $|E|$ – šio tinklo briaunų skaičių, $\overline{d}$ – vidutinį viršūnės laipsnį, WCC (%) ir SCC (%) – dalį viršūnių, patenkančių į didžiausią silpnai (WCC) ir stipriai (SCC) jungią komponentę procentais. Laiko požymis indikuoja, ar tinklas turi evoliucijos bėgant laikui informaciją.

Tinklas	$\|V\|$	$\|E\|$	$\overline{d}$	WCC (%)	SCC (%)	Laiko požymis
WIK (Wikipedia Vote)	7115	103689	29,15	99,31	18,27	-
SCC (Statisticians' Coauthorship)	2693	21603	16,04	98,55	42,33	-
SPA (StackExchange Physics Answers)	48445	151458	6,25	98,04	6,80	+
SPC (StackExchange Physics Comments)	45541	239533	10,52	99,58	38,76	+
SSA (StackExchange Statistics Answers)	49869	109508	4,39	95,13	4,14	+
SSC (StackExchange Statistics Comments)	53462	183295	6,86	98,83	32,55	+

Kaip ir būdinga daugumai realaus pasaulio tinklų, nagrinėjamuose tinkluose pastebimas laipsninis viršūnių laipsnių skirstinys. Diagramose žemiau pateikiama tinklo *SPA* viršūnių įėjimo ir išėjimo laipsnių skirstiniai (žr. 7 pav.). Visų kitų darbui sumodeliuotų tinklų (*SPC*, *SSA* ir *SSC*) viršūnių laipsnių skirstiniai pateikiami priede Nr. 2.



(a) (b)

7 pav. Viršūnių laipsnių skirstiniai tinkle *SPA*. Duomenis abejose taškinėse diagramose konvertavus į logaritmo pagrindu dešimt skalę matoma į tiesę panaši priklausomybė. Tokia tendencija yra būdinga laipsniniam skirstiniui.

5. Eksperimentai

Šioje darbo dalyje atliekami ankstesniuose skyriuose nagrinėtų jungčių prognozės metodų, paremtų diklikos klasterizacijos koeficientu, tyrimai. Bandymai atliekami duomenų rikimo skyriuje aprašytuose tinkluose.

Šiame skyriuje pirmiausia nagrinėjami klasterizacijos koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ įvertinimo algoritmai bei šių koeficientų priklausomybė nuo jų parametrų k, w, t. Vėliau kiekvienas koeficientas išmėginamas prognozuojant jungtis. Galiausiai, analizuojami gauti rezultatai.

5.1. Koeficientų apskaičiavimas

Šiame poskyryje aprašomi galimi koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ įvertinimo algoritmai, pagal lygybes, apibrėžtas diklikos klasterizacijos koeficiento praplėtimų poskyryje (žr. 29, 30 ir 31 lygybes atitinkamai).

Algoritmų aprašuose naudojamasi šiame darbe taikomais grafų ir aibių teorijų žymėjimais. Kiekvienas algoritmas turi tam tikrą įvestį, išvestį bei gali būti parametrizuojamas.

Algoritmas 1: $C^{IN}(D,k)$ įvertinimo algoritmas orientuotam tinklui $D(V,E)$.

Įvestis : $D(V,E)$
Parametras: k
Išvestis : $C^{IN}(D,k)$

$N_1 \leftarrow 0$
$N_2 \leftarrow 0$
for $u \in V$ **do**
  for $v \in N_{out}(u)$ **do**
    for $x \in N_{in}(v) \setminus \{u\}$ **do**
      for $y \in N_{out}(x) \setminus \{u,v\}$ **do**
        if $d_{in}(v) = k$ **then**
          if $(u,y) \in E$ **then**
            $N_1 \leftarrow N_1 + 1$
          else
            $N_2 \leftarrow N_2 + 1$
          end
        end
      end
    end
  end
end
$C^{IN}(D,k) \leftarrow N_1/(N_1 + N_2)$

Koeficiento $C^{IN}(D,k)$ įvertinimo algoritmas orientuotam tinklui $D(V,E)$ pateikiamas pse-

udokodu aukščiau. $N_{out}(v)$ žymi viršūnių, į kurias viršūnė v turi orientuotą briauną, aibę. Formaliai $N_{out}(v)$ yra funkcija, kuri viršūnei $v \in V$ gražina viršūnių sąrašą $\{x : (v,x) \in E\}$.

Koeficiento $C^W(D,w)$ įvertinimas apima du žingsnius. Pirmiausia sudaromas sąrašas W viršūnių liudininkių kiekvienai tiriamai briaunai $i_2 \to i_4$ (remiantis tinklo motyvu iš 5 pav.). $N_{in}(v)$ žymi viršūnių, kurios turi orientuotą briauną į viršūnę v, aibę. $W(u,v)$ yra aibė viršūnių, kurios liudija tam tikrą galimą jungtį $(u,v), u,v \in V$. W čia taip pat yra funkcija, kurios apibrėžimo sritis $dom(W)$ yra poros $(u,v), u, v \in V$.

Algoritmas 2: Liudininkų sąrašo W tinklui $D(V,E)$ generavimo algoritmas.

Įvestis : $D(V,E)$
Išvestis: W
$W \leftarrow \varnothing$
for $u \in V$ **do**
  for $v \in N_{out}(u)$ **do**
    for $x \in N_{in}(v) \setminus \{u\}$ **do**
      for $y \in N_{out}(x) \setminus \{u,v\}$ **do**
        if $W(u,y) = \varnothing$ **then**
          $W(u,y) \leftarrow \{x\}$ // x liudija ryšį (u,y)
        else
          $W(u,y) \leftarrow W(u,y) \cup \{x\}$ // x liudija ryšį (u,y)
        end
      end
    end
  end
end

Algoritmas 3: $C^W(D,w)$ įvertinimo algoritmas orientuotam tinklui $D(V,E)$.

Įvestis : $D(V,E), W$
Parametras: w
Išvestis : $C^W(D,w)$
$W_1 \leftarrow 0$
$W_2 \leftarrow 0$
for $(u,v) \in dom(W)$ **do**
  if $|W(u,v)| = w$ **then**
    if $(u,v) \in E$ **then**
      $W_1 \leftarrow W_1 + 1$
    else
      $W_2 \leftarrow W_2 + 1$
    end
  end
end
$C^W(D,w) \leftarrow W_1/(W_1 + W_2)$

Antras žingsnis yra koeficiento $C^W(D,w)$ parametrui w apskaičiavimas. Tai daroma įvertinus,

kiek viršūnių porų $(u,v) \in dom(W)$ iš tiesų suformuoja briauną tinkle $D(V,E)$, su sąlyga, kad liudininkų yra lygiai w ($|W(u,v)| = w$) (žr. 3 alg.).

Koeficiento $C^T(D,t)$ įvertinimas taip pat apima du žingsnius ir iš esmės yra labai panašus į $C^W(D,w)$ atvejį. Pirmame žingsnyje yra sudaromas sąrašas T viršūnių i_1 ir i_2 bendrai sekamų viršūnių (bendrų interesų) (remiantis tinklo motyvu iš 5 pav.). Panašiai kaip ir ankstesniuose atvejuose, T čia yra funkcija, kurios apibrėžimo sritis, žymima $dom(T)$, yra viršūnių poros $(u,v), u, v \in V$.

Algoritmas 4: Bendrų interesų sąrašo T tinklui $D(V,E)$ generavimo algoritmas.

```
   Įvestis : D(V,E)
   Išvestis: T
1  T ← ∅
2  for u ∈ V do
3  |  for v ∈ N_out(u) do
4  |  |  for x ∈ N_in(v) \ {u} do
5  |  |  |  for y ∈ N_out(x) \ {u,v} do
6  |  |  |  |  if T(u,y) = ∅ then
7  |  |  |  |  |  T(u,y) ← {v}                    // u ir x bendrai seka v
8  |  |  |  |  else
9  |  |  |  |  |  T(u,y) ← T(u,y) ∪ {v}           // u ir x bendrai seka v
10 |  |  |  |  end
11 |  |  |  end
12 |  |  end
13 |  end
14 end
```

Algoritmas 5: $C^T(D,t)$ įvertinimo algoritmas orientuotam tinklui $D(V,E)$.

```
   Įvestis     : D(V,E), T
   Parametras: t
   Išvestis    : C^T(D,t)
1  T_1 ← 0
2  T_2 ← 0
3  for (u,v) ∈ dom(T) do
4  |  if |T(u,v)| = t then
5  |  |  if (u,v) ∈ E then
6  |  |  |  T_1 ← T_1 + 1
7  |  |  else
8  |  |  |  T_2 ← T_2 + 1
9  |  |  end
10 |  end
11 end
12 C^T(D,t) ← T_1/(T_1 + T_2)
```

Antrame žingsnyje vyksta dydžio $C^T(D,t)$ parametrui t apskaičiavimas. Tai atliekama įver-

tinus, kiek viršūnių porų $(u,v) \in dom(T)$ iš tiesų suformuoja briaunas tinkle $D(V,E)$, su sąlyga, kad bendrai sekamų viršūnių yra lygiai t ($|T(u,v)| = t$) (žr. 5 alg.).

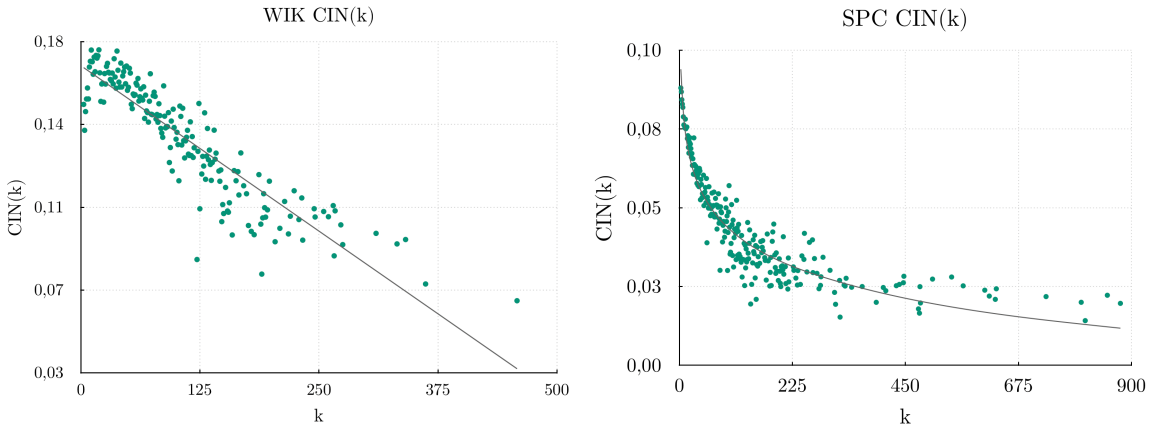
Visi penki algoritmai tyrimo metu įgyvendinti *Go* programavimo aplinkoje (žr. priedą Nr. 1). Labai svarbu paminėti, jog aukščiau aprašyti algoritmai yra abstraktūs ir suprastinti lengvesniam skaitovo suvokimui. Tai reiškia, jog jie neapima konkrečių duomenų struktūrų naudojimo, optimizacijų ir skaičiavimų išlygiagretinimo, kurie buvo taikyti įgyvendinant algoritmus.

5.2. Koeficientų priklausomybė nuo parametrų

Atlikus koeficientų matavimus realiuose tinkluose, naudojantis poskyryje aukščiau aprašytais algoritmais, pastebėtos ganėtinai ryškios šių koeficientų įverčių priklausomybės nuo jų parametrų. Matavimai atlikti visuose tiriamuose tinkluose (tinklo briaunų ir viršūnių skaičius nemažintas). Gauti rezultatai su komentarais pateikiami žemiau. Rezultatai iliustruojami taškinėmis diagramomis, kurių horizontaliose ašyse vaizduojami parametrai k, w ir t, o vertikaliose ašyse – atitinkamai koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ įverčiai šiems parametrams. Diagramose taip pat galima stebėti, kiek padidėja briaunos egzistavimo tikimybė esant vienokiai ar kitokiai parametro reikšmei, kai briaunos tikimybė orientuotame tinkle yra lygi $\frac{|E|}{|V| \cdot (|V|-1)}$ (žr. 9 lentelę).

9 lentelė. Apytikslės briaunos egzistavimo tikimybės (p) tiriamuose tinkluose.

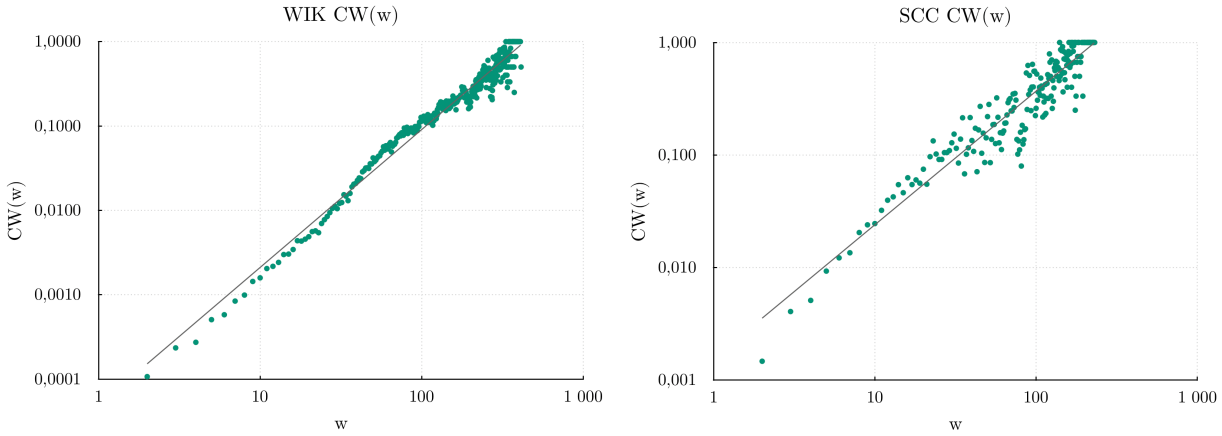
Tinklas	**WIK**	**SCC**	**SPA**	**SPC**	**SSA**	**SSC**
$\|V\|$	7115	2693	48445	45541	49869	53462
$\|E\|$	103689	21603	151458	239533	109508	183295
p	$2{,}05 \cdot 10^{-3}$	$2{,}98 \cdot 10^{-3}$	$6{,}45 \cdot 10^{-5}$	$1{,}16 \cdot 10^{-4}$	$4{,}4 \cdot 10^{-5}$	$6{,}41 \cdot 10^{-5}$



(a) (b)

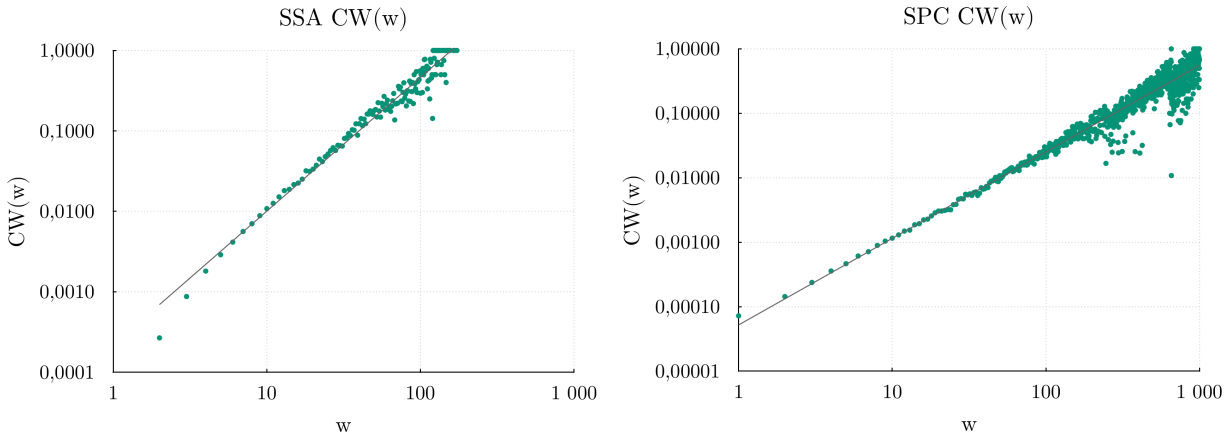
8 pav. Koeficiento $C^{IN}(D,k)$ priklausomybė nuo parametro k. Taškinės diagramos (a) ir (b) vaizduoja šią priklausomybę tinkluose *WIK* ir *SPC* atitinkamai. Visuose tiriamuose tinkluose nesunku pastebėti ganėtinai ryškų koeficiento mažėjimą augant dydžiui k (žr. priedą Nr. 3).

Kaip galima pastebėti iš diagramų pateikiamų aukščiau ir prieduose, koeficiento $C^{IN}(D,k)$ reikšmė ganėtinai akivaizdžiai priklauso nuo parametro k reikšmės. Kuo ši reikšmė didesnė, tuo mažesnis koeficiento įvertis. Kita vertus, nors tendencija ir gana aiški, koeficiento reikšmė visame parametro k matavimo intervale daugumoje tinklų pakinta ganėtinai nedideliu įverčiu (2 % – 5 %). Tokio nedidelio pokyčio rezultatas gali būti ne itin kokybiška prognozė, kadangi rezultatuose yra ganėtinai sunku įžvelgti pakankamai ryškius jungties tikėtinumo skirtumus kintant parametrui k.



(a) (b)

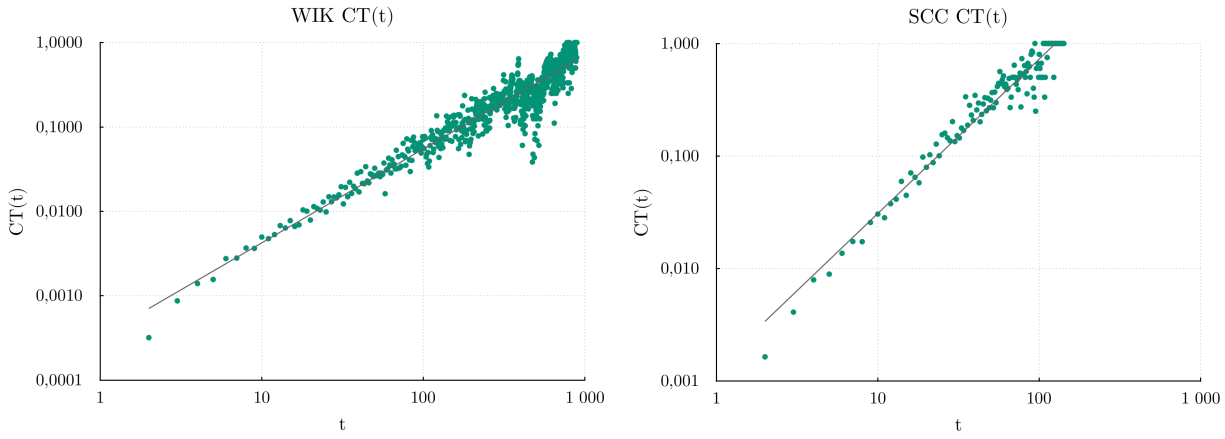
9 pav. Koeficiento $C^{W}(D,w)$ priklausomybė nuo parametro w. Taškinės diagramos (a) ir (b) vaizduoja šią priklausomybę tinkluose *WIK* ir *SCC* atitinkamai. Diagramose abi ašys paverstos į logaritminę skalę pagrindu dešimt. Nesunku pastebėti ganėtinai ryškų koeficiento didėjimą augant dydžiui w. Taip pat svarbu pastebėti, kad koeficiento vertė pakinta praktiškai visame savo galimų reikšmių intervale.



(a) (b)

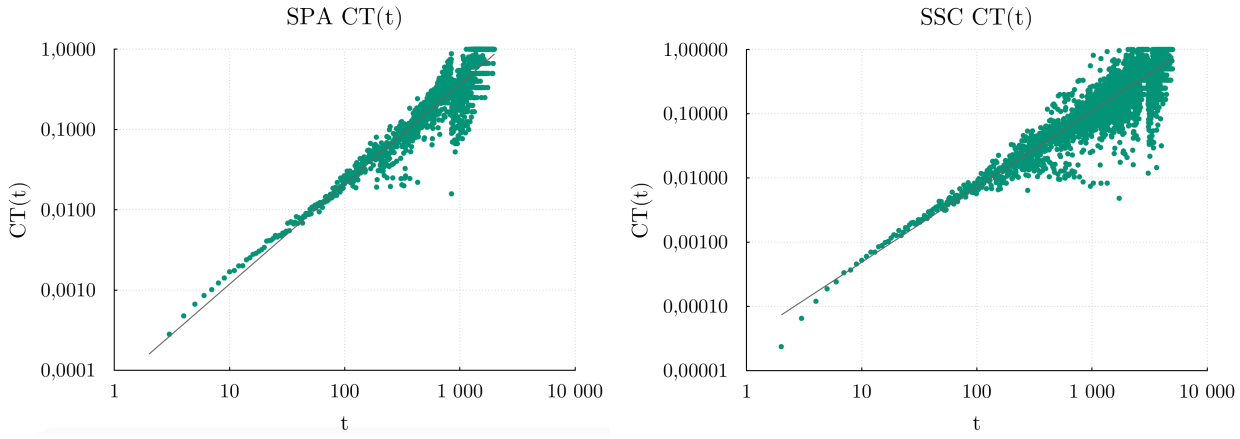
10 pav. Koeficiento $C^{W}(D,w)$ priklausomybė nuo parametro w. Taškinės diagramos (a) ir (b) vaizduoja šią priklausomybę tinkluose *SSA* ir *SPC* atitinkamai. Diagramose abi ašys paverstos į logaritminę skalę pagrindu dešimt. Šiuose tinkluose taip pat stebimas ryškus koeficiento reikšmės augimas augant parametrui w, tačiau tinkle *SPC* matomas kiek lėtesnis augimas.

Iš diagramų aukščiau galima pastebėti ryškią koeficiento $C^W(D,w)$ reikšmės priklausomybę nuo parametro w. Kuo didesnė šio parametro reikšmė, tuo didesnis koeficiento įvertis. Be to, stebimame intervale nagrinėjamas koeficientas pakinta praktiškai visame savo reikšmių intervale (nuo minimalios iki maksimalios reikšmės). Tokia tendencija stebima visuose tiriamuose tinkluose (žr. priedą Nr. 3).



(a) (b)

11 pav. Koeficiento $C^T(D,t)$ priklausomybė nuo parametro t. Taškinė diagrama (a) vaizduoja šią priklausomybę tinkle *WIK*, o diagrama (b) – tinkle *SCC*. Abiejų diagramų abi ašys paverstos į logaritminę skalę pagrindu dešimt. Tinkluose šis koeficientas elgiasi panašiai kaip ir koeficientas $C^W(D,w)$, tik daugumoje jų auga kiek lėčiau augant parametrui t.



(a) (b)

12 pav. Koeficiento $C^T(D,t)$ priklausomybė nuo parametro t. Taškinės diagramos (a) ir (b) vaizduoja vaizduoja šią priklausomybę tinkluose *SPA* ir *SSC* atitinkamai. Abiejų diagramų abi ašys paverstos į logaritminę skalę pagrindu dešimt. Koeficiento didėjimo augant parametrui t tendencija stebima visuose tiriamuose tinkluose (žr. priedą Nr. 3).

Apibendrindami galime pastebėti, kad koeficientų realiuose tinkluose matavimų rezultatai, išanalizuoti šiame poskyryje, indikuoja ganėtinai aiškią kiekvieno tiriamo koeficiento reikšmių

priklausomybę nuo su juo susijusio parametro reikšmių. Tai mums yra itin reikšmingas rezultatas, kadangi šis faktas supaprastina jungčių prognozės procesą, kaip parodysime tolimesniuose eksperimentuose.

5.3. Prognozės duomenų paruošimas

Šiame darbe tiriami klasterizacijos koeficientai išbandomi sprendžiant jungčių prognozės uždavinius duomenų surinkimo skyriuje aprašytuose tinkluose. Pirmiausia tyrimai atliekami tinkluose be laiko požymio, mūsų pavadintuose *WIK* ir *SCC*. Dėl briaunų susiformavimo laiko nežinojimo šiuose tinkluose atliekama struktūrinė jungčių prognozė. Likę tyrimai organizuojami tinkluose, mūsų pavadintuose *SPA*, *SPC*, *SSA* ir *SSC*. Visi šie tinklai turi laiko požymį, todėl jiems atliekama laikinė jungčių prognozė. Svarbu pastebėti, kad tiriami tinklai yra ganėtinai skirtingi dydžio, tankumo, srities (socialiniai, informaciniai) ir nagrinėjamų koeficientų įverčių prasmėmis.

Dėl didelės tinklų apimties darbe atliekama seka tam tikrų tinklo transformacijų. Šios transformacijos reikalingos tam, kad tyrimo duomenys būtų tinkamesni bandymams ir turimiems skaičiavimo resursams. Duomenų apdorojimas atliekamas trimis žemiau aprašomais etapais.

1. *Tinklo sumažinimas*. Tinklas sumažinamas pirmiausia atsitiktinai pasirenkant $8000 - 15000$ viršūnių. Jeigu tinklas yra mažesnis nei nurodyti rėžiai, mažinimas nėra atliekamas. Šis žingsnis yra reikalingas tam, kad su turimais skaičiavimo resursais prognozės uždavinys būtų išsprendžiamas su normaliomis laiko sąnaudomis. Viršūnių skaičius parenkamas atsižvelgus į tinklo tankumą: jei tinkle yra daug briaunų, parenkamas mažesnis viršūnių skaičius.

2. *Tinklo padalijimas į mokymo ir testavimo potinklius*. Jei tinklas turi laiko požymį, jo briaunos padalijamos į dvi nepersikertančias aibes, kur testavimo aibė turi $10\,\%$ naujausių briaunų, o mokymo aibė turi likusias $90\,\%$ briaunų. Jei tinklas neturi laiko požymio, į mokymo ir testavimo aibes jis padalinamas atsitiktinai ir nepriklausomai tuo pat santykiu.

3. *Jungios komponentės parinkimas*. Tinklo dalijimas į du potinklius gali izoliuoti kai kurias viršūnes ar klikas. To rezultatas gali būti nereikalingas triukšmas prognozės metodo kokybės matavimuose. Šis triukšmas gali būti sumažinamas nuo mokymo potinklio atskiriant didžiausią silpnai jungią komponentę (angl. WCC, Weakly Connected Component) [GMC+15]. Ši komponentė tuomet imama kaip galutinis mokymo potinklis. Galutinis testavimo potinklis yra gaunamas suvienodinus šio potinklio viršūnių aibę su galutinio mokymo potinklio viršūnių aibe.

10 lentelė. Prognozei paruoštų duomenų rinkinių informacija. $|V_M|$ ir $|V_T|$ atitinkamai žymi mokymo ir testavimo potinklių viršūnių skaičius, o $|E_M|$ ir $|E_T|$ – šių potinklių briaunų skaičius.

Tinklas	Mokymo laiko rėžiai	Testavimo laiko rėžiai	$\|V_M\|$	$\|V_T\|$	$\|E_M\|$	$\|E_T\|$
WIK	-	-	6827	3332	93296	10125
SCC	-	-	2643	1540	19410	2144
SPA	2010-11-03 - 2017-08-15	2017-08-15 - 2018-09-02	6779	601	13309	977
SPC	2010-11-02 - 2017-12-06	2017-12-07 - 2018-09-01	8718	834	25877	1888
SSA	2010-07-19 - 2016-03-20	2016-03-21 - 2018-08-30	6154	632	10613	909
SSC	2010-07-19 - 2016-07-29	2016-07-29 - 2018-08-31	5809	762	11499	1386

5.4. Jungčių prognozė

Jungčių prognozė atliekama mokymo potinklyje nesančias jungtis įvertinant tam tikru jungties tikėtinumo įverčiu. Tuomet šios potencialios jungtys yra surikiuojamos pagal priskirtą tikėtinumo įvertį nuo didžiausios iki mažiausios jo reikšmės. Galiausiai, prognozės rezultatu yra skelbiamas tam tikras skaičius k didžiausius įverčius turinčių jungčių. Šiame tyrime išmėginami visais trim nagrinėjamais klasterizacijos koeficientais paremti jungties tikėtinumo indeksai: *įėjimo laipsnių agregacijos* indeksas s_{xy}^{CIN}, *liudininkų* indeksas s_{xy}^{CW} bei *bendrų interesų* indeksas s_{xy}^{CT} (žr. 32, 33 ir 34 išraiškas atitinkamai).

Kaip galima pastebėti iš tiriamų jungčių tikėtinumo indeksų apibrėžimų, jų įvertinimas yra ganėtinai sudėtingas procesas, skaičiavimo resursų prasme (kiekvienai viršūnių porai (x, y) reikia skaičiuoti atitinkamą koeficientą, priklausomai nuo koeficiento parametro). Šį skaičiavimų etapą galima nesudėtingai supaprastinti dviem būdais:

1. Pirmas būdas yra naudotis tinklo statistikomis. Tai reiškia, jog visų pirma reikia apskaičiuoti atitinkamą parametrą (k, w arba t) vertinamai jungčiai (x, y). Tuomet apytikrį tikėtinumo įvertį šiai jungčiai parinkti naudojantis iš anksto paruošta koeficiento priklausomybės nuo atitinkamo parametro statistika (pavyzdžiui, viena koeficientų priklausomybės nuo parametrų poskyryje nagrinėjamų taškinių diagramų).

2. Antras (dar paprastesnis) būdas yra naudotis tik koeficiento parametrais. Kaip parodėme koeficientų priklausomybės nuo parametrų poskyryje, kiekvienas iš nagrinėjamų koeficientų turi ganėtinai aiškią didėjimo ar mažėjimo tendenciją kintant atitinkamam parametrui. Tai reiškia, kad ir kiekvienas iš šiais koeficientais paremtų jungčių tikėtinumo indeksų turės tą pačią tendenciją (tikėtinumo indeksas yra tiesiog atitinkamo koeficiento įvertis priklausomai nuo atitinkamo parametro, apskaičiuojamo porai (x, y)). Iš visų šių faktų išplaukia

tai, kad prognozės procesą supaprastinti galima jungties tikėtinumo įverčiu tiesiog pasirenkant atitinkamo parametro (k, w arba t) įvertį šiai jungčiai ir surikiuoti gautą potencialių jungčių sąrašą (pagal parametrą) didėjimo arba mažėjimo tvarka, priklausomai nuo to, kokia tendencija kinta atitinkamas koeficientas kintant jo parametrui.

Šiame darbe prognozės procesui supaprastinti pasirenkamas antrasis metodas, kadangi prognozės uždavinio sprendimui tikimybinis jungties tikėtinumo įvertis, teikiamas pirmojo metodo, nėra būtinas.

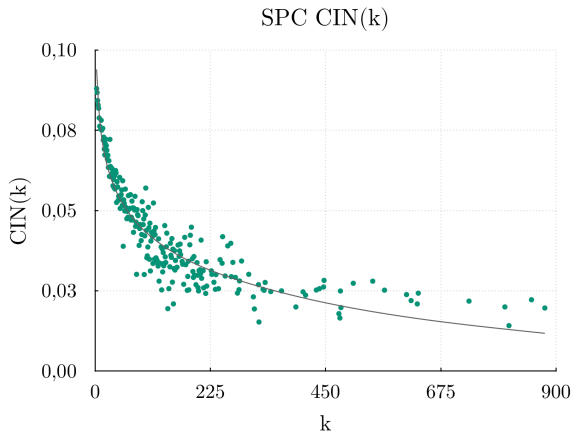
- Įėjimo laipsnių agregacijos indekso atveju tikėtinumo įverčiu vertinamai porai (x, y) yra parenkama parametro k įverčių suma (agregacija) vietoje koeficiento $C^{IN}(D,k)$ įverčių sumos (žr. 32 išraišką). Kadangi pastarasis koeficientas mažėja didėjant parametrui k, gautas potencialių jungčių sąrašas surikiuojamas didėjimo tvarka (mažesnius įverčius turinčios jungtys yra skelbiamos labiau tikėtinomis).

- Liudininkų indekso atveju jungties tikėtinumo įverčiu yra parenkamas jungties tarp poros (x, y) liudininkų skaičius w. Kadangi koeficientas $C^{W}(D,w)$ tiriamuose tinkluose didėja augant parametrui w, gautas potencialių jungčių sąrašas surikiuojamas mažėjimo tvarka (didesnius įverčius turinčios jungtys yra skelbiamos labiau tikėtinomis).

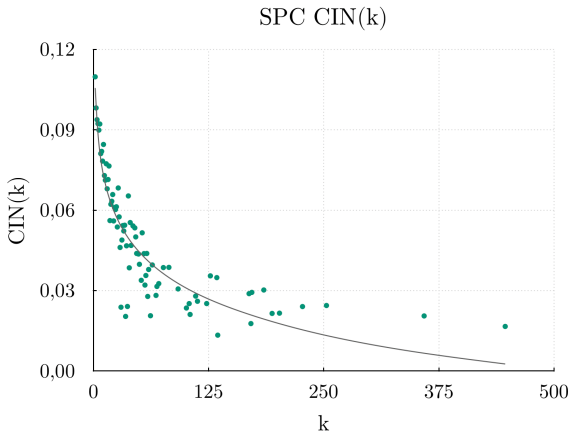
- Bendrų interesų indekso atvejis yra visiškai analogiškas liudininkų indeksui. Jungties tikėtinumo įverčiu yra parenkamas parametras t. Kadangi koeficiento įvertis $C^{T}(D,t)$ auga didėjant parametrui t, gautas potencialių jungčių sąrašas surikiuojamas mažėjimo tvarka.

Kiekvieno iš nagrinėjamų jungčių tikėtinumo indeksų prognozės kokybė yra įvertinama prognozės metodų vertinimo poskyryje nagrinėjamomis metrikomis: ROC ir jautrumo-specifiškumo kreivėmis bei plotais po jomis (AUROC ir AUPR atitinkamai). Minėtosios kokybės kreivės gaunamos įvertinant prognozės rezultatų kokybę iteratyviai. Tai reiškia, jog kiekvienos iteracijos metu yra didinamas teigiamos prognozės rezultatų skaičius k. Kitaip tariant, kiekvienoje tokioje iteracijoje vis daugiau potencialių jungčių iš pagal tikėtinumo indeksą surikiuoto sąrašo yra skelbiamos prognozės rezultatais. Šių prognozės rezultatų kokybė yra patikrinama su testavimo potinkliu. Tai padaroma įvertinant kokybės kreivių sudarymui reikalingas metrikas, tokias kaip teisingai teigiamų ir klaidingai teigiamų prognozių rodikliai, jautrumas bei specifiškumas.

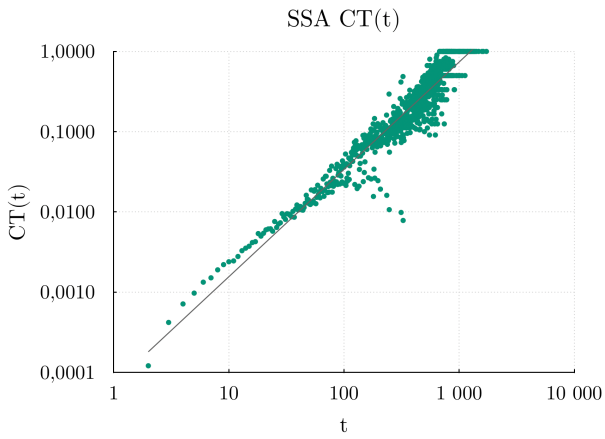
Toliau šiame poskyryje yra nagrinėjami jungties tikėtinumo indeksų indeksų kokybės matavimų rezultatai mokymo ir testavimo duomenų rinkiniams, parengtiems pagal prognozės duomenų paruošimo poskyryje pateikiamas instrukcijas iš tinklų, aprašytų duomenų surinkimo skyriuje (*WIK*, *SCC*, *SPA*, *SPC*, *SSA*, *SSC*).



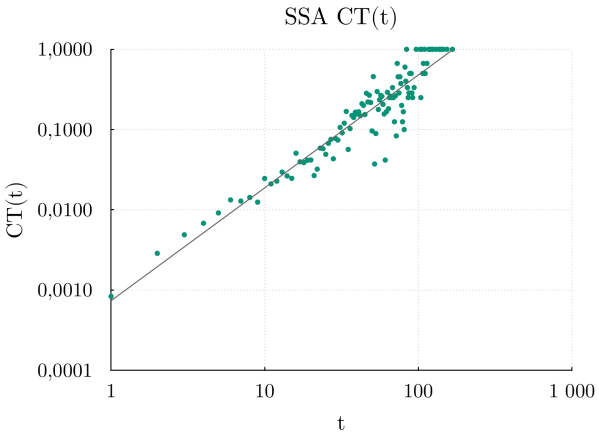
(a) Originalus (nesumažintas) tinklas



(b) Mokymo potinklis (sumažintas tinklas)



(c) Originalus (nesumažintas) tinklas



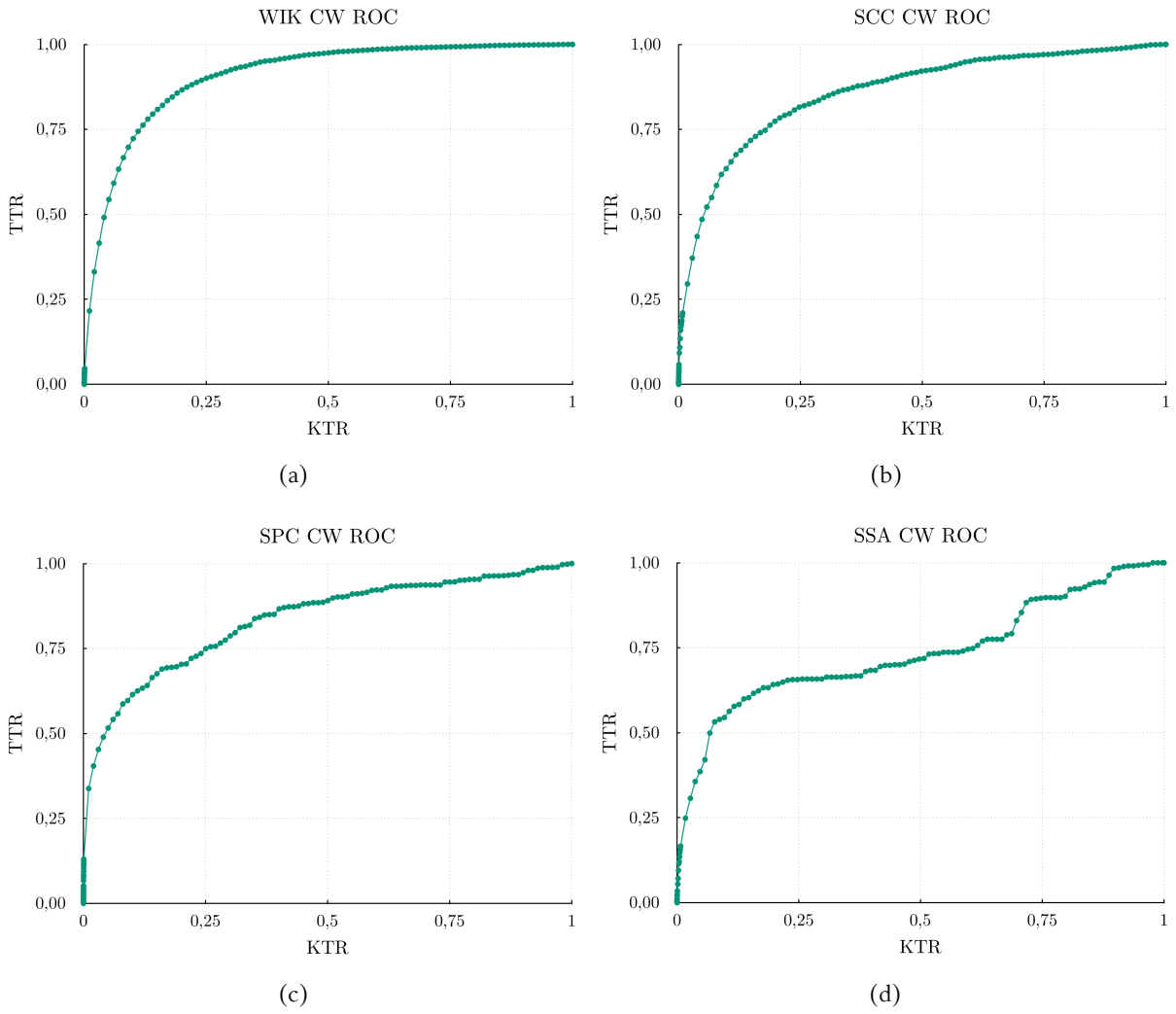
(d) Mokymo potinklis (sumažintas tinklas)

13 pav. Mokymo potinkliuose tiriamų koeficientų priklausomybės nuo jų parametrų kitimo tendencija iš esmės nekinta lyginant su originaliais tinklais. Visuose mokymo potinkliuose matomas ganėtinai ryškus kiekvieno koeficiento didėjimas arba mažėjimas priklausomai nuo jo parametro.

Kaip ir tikėtasi, visų trijų tiriamų klasterizacijos koeficientų kitimo tendencijos kintant jų parametrų reikšmėms mokymo potinkliuose išlieka nepakitusios lyginant su originaliais tinklais, iš kurių šie potinkliai gauti (žr. 13 pav.). Tai yra svarbi sąlyga priimant tolimesnius sprendimus jungčių prognozės metu, kadangi rikiuodami jungčių tikėtinumo įverčius remiamės koeficientų parametrų tendencijomis. Detalios diagramos kiekvienam tinklui pateikiamos priede Nr. 4.

5.4.1. Liudininkų indekso prognozės kokybė

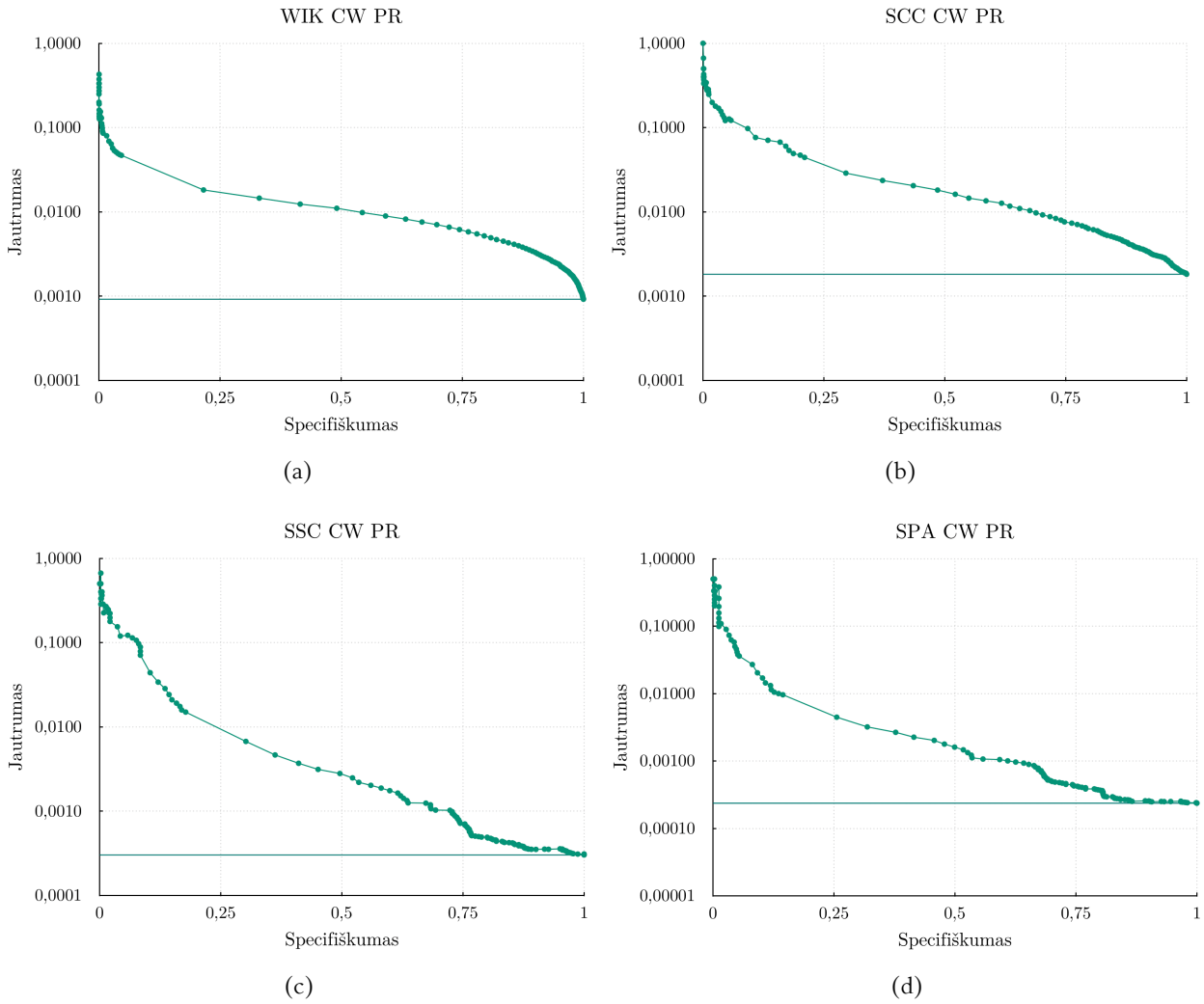
Liudininkų indekso prognozės kokybės tyrimo rezultatai indikuoja ganėtinai kokybiškus rezultatus visuose tiriamuose tinkluose. ROC kreivės yra aiškiai nutolusios nuo atsitiktinės prognozės (įstrižainės ROC erdvėje) (žr. 14 pav.). Atitinkamai ir AUROC įverčiai indikuoja kokybišką prognozę: visuose tiriamuose tinkluose išskyrus *SPA* ir *SSA* jis yra apie 0,8 arba daugiau. Tinkluose *SPA* ir *SSA* šis dydis yra apie 0,75 (žr. 11 lentelę).



14 pav. ROC kreivės, gautos vertinant liudininkų indekso s_{xy}^{CW} prognozės kokybę. KTR atspindi klaidingai teigiamų prognozių rodiklį, TTR – teisingai teigiamų. ROC kreivės visiems tiriamiems tinklams pateikiamos priede Nr. 5.

Jautrumo-specifiškumo kreivės (angl. precision-recall arba PR) taip pat rodo neblogą liudininkų indekso prognozės kokybę. Kaip matyti iš 15 pav., esant nedidelėms specifiškumo reikšmėms, jautrumas pasiekia labai aukštus įverčius. Tai liudija faktą, kad, kai prognozės rezultatu yra skelbiamas nedidelis kiekis visų potencialių jungčių, nemaža dalis jų iš tiesų egzistuoja testavimo potinklyje. Taip pat galime pastebėti ir tai, kad jautrumo-specifiškumo kreivės tik pačiame specifiškumo skalės gale (specifiškumui esant prie vieneto) pasiekia atsitiktinės prognozės įverčius (horizontali linija diagramose žemiau). Nepaisant to, jautrumo įverčių nykimas didinant prognozės rezultatu skelbiamų potencialių jungčių kiekį nyksta labai sparčiai.

Panašius rezultatus sufleruoja ir AUPR įverčiai, kurie tiriamuose tinkluose yra gerokai didesni nei atsitiktinės prognozės atveju. Pavyzdžiui tinkle *SPC* AUPR yra lygus maždaug 0,02, kai atsitiktinės prognozės atveju šis dydis yra artimas 10^{-4} (200 kartų mažesnis). Visi AUPR įverčiai liudininkų indeksui tiriamuose duomenų rinkiniuose pateikiami 12 lentelėje.

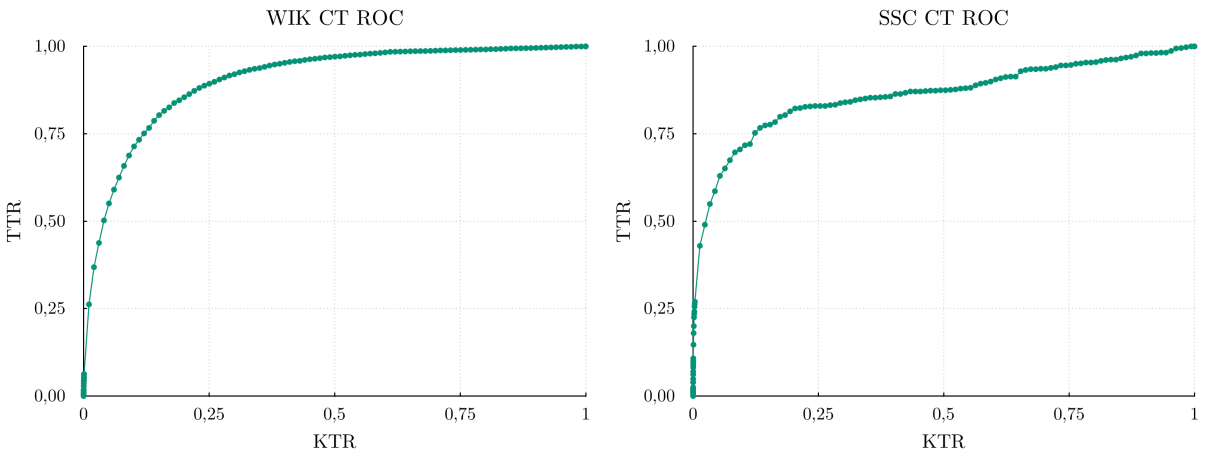


15 pav. Jautrumo-specifiškumo kreivės, gautos vertinant liudininkų indekso s_{xy}^{CW} prognozės kokybę. Kreivių jautrumo ašys yra pateikiamos logaritmo pagrindu dešimt skalėje norint geriau atskleisti jų elgesį esant mažoms jautrumo reikšmėms. Jautrumo-specifiškumo kreivės visiems tiriamiems tinklams pateikiamos priede Nr. 5.

5.4.2. Bendrų interesų indekso prognozės kokybė

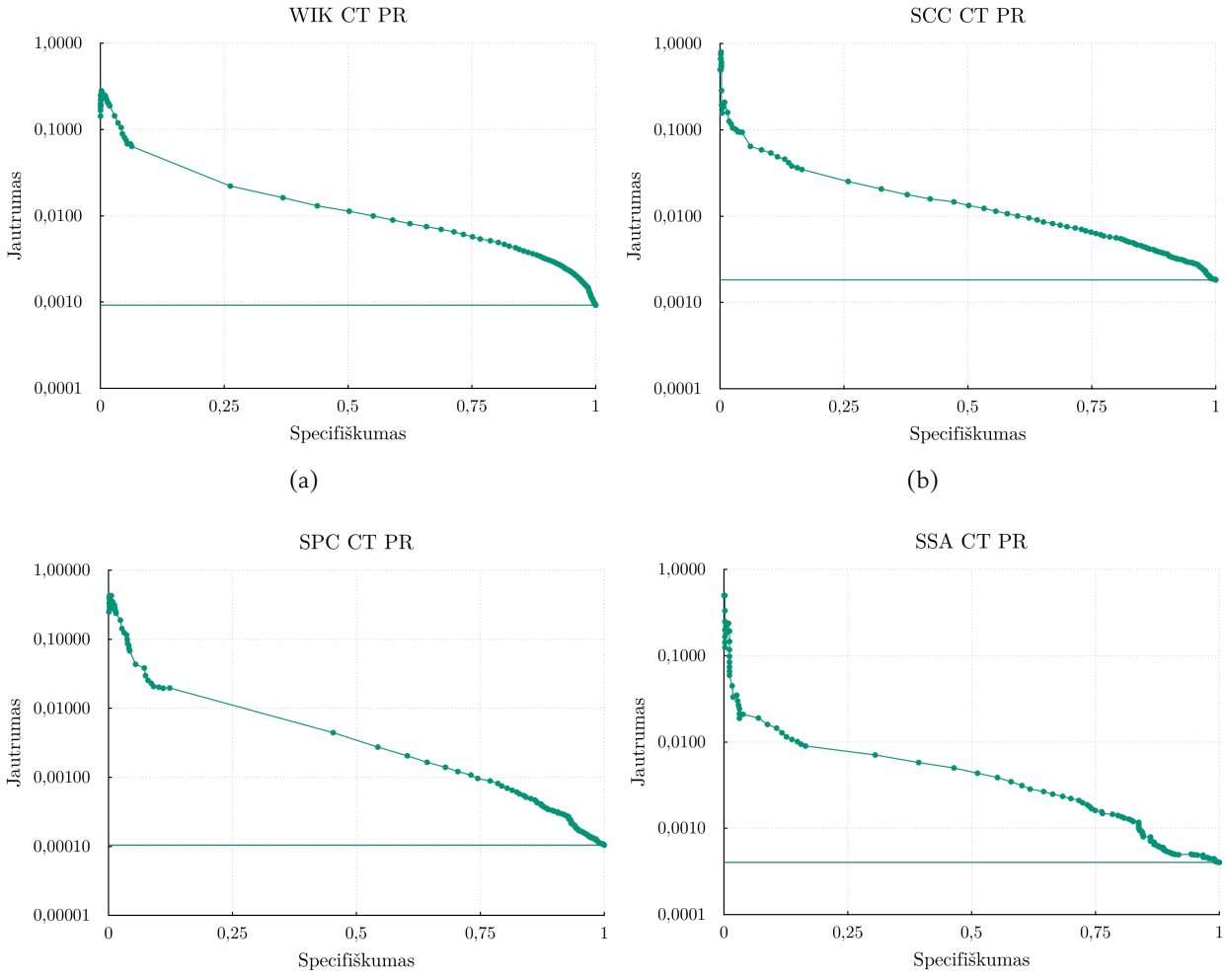
Bendrų interesų indekso s_{xy}^{CT} prognozės kokybės tyrimo metu gauti į liudininkų indekso tyrimo atvejį panašūs ir netgi kokybiškesni rezultatai. ROC kreivės ir čia yra ryškiai nutolusios nuo atsitiktinę prognozę indikuojančios įstrižainės (žr. 16 pav.), o AUROC įvertis visuose tirtuose tinkluose yra didesnis arba lygus 0,84 (žr. 11 lentelę).

Žvelgiant į jautrumo-specifiškumo kreives (žr. 17 pav.), taip pat matomi į liudininkų indekso atvejį panašūs rezultatai. Čia esant nedidelėms specifiškumo reikšmėms jautrumas taip pat pasiekia gana aukštus (nors ir kiek mažesnius nei liudininkų indekso atveju) įverčius. Kaip ir liudininkų indekso atveju, jautrumas čia irgi gana sparčiai nyksta augant specifiškumo reikšmėms, o stebimi AUPR įverčiai yra gerokai didesni nei būdinga atsitiktinei prognozei (žr. 12 lentelę).



(a) (b)

16 pav. ROC kreivės, gautos vertinant bendrų interesų indekso s_{xy}^{CT} prognozės kokybę. ROC kreivės visiems tiriamiems tinklams pateikiamos priede Nr. 5.



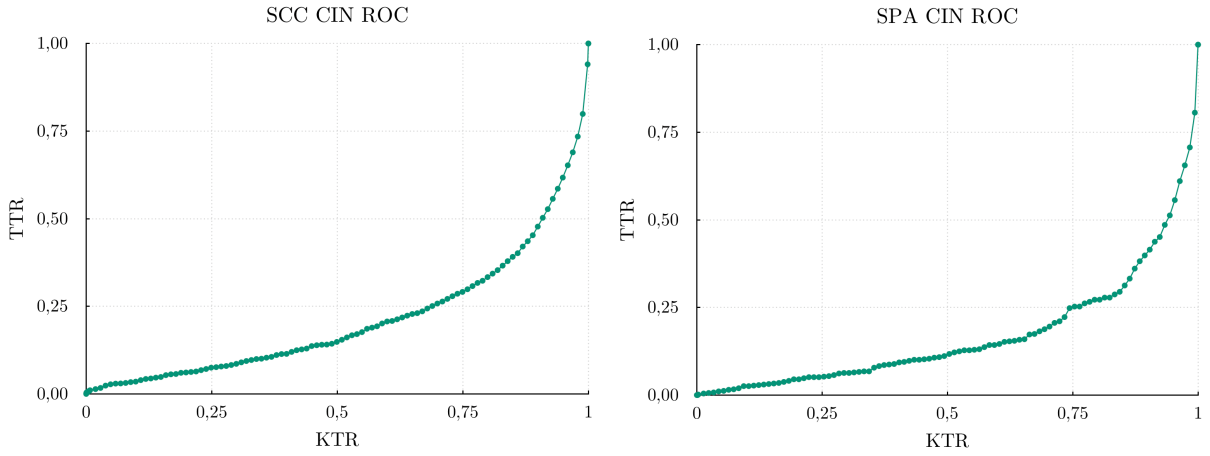
(a) (b)

(c) (d)

17 pav. Jautrumo-specifiškumo kreivės, gautos vertinant bendrų interesų indekso s_{xy}^{CT} prognozės kokybę. Kreivių jautrumo ašys yra pateikiamos logaritmo pagrindu dešimt skalėje norint geriau atskleisti jų elgesį esant mažoms jautrumo reikšmėms. Jautrumo-specifiškumo kreivės visiems tiriamiems tinklams pateikiamos priede Nr. 5.

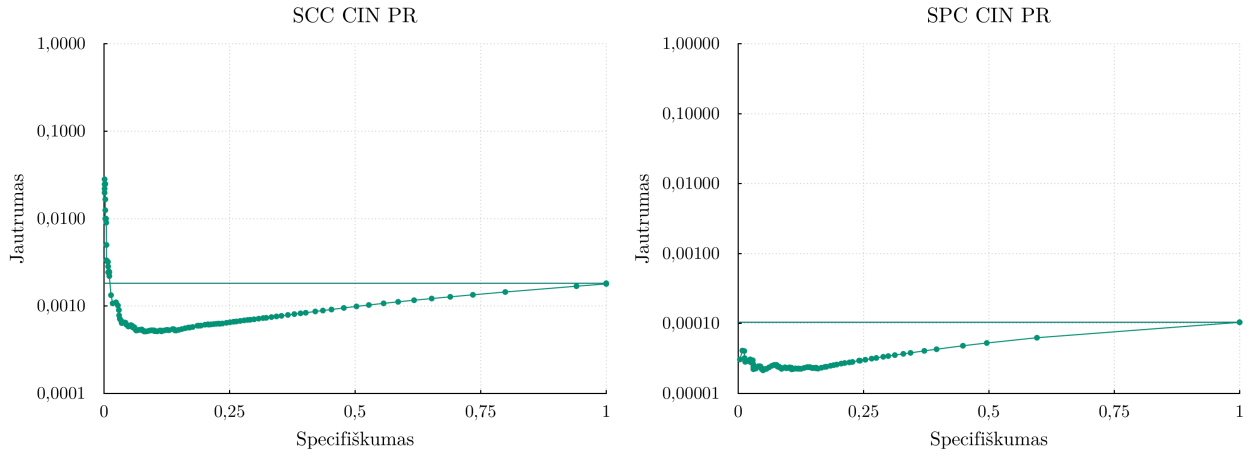
5.4.3. Įėjimo laipsnių agregacijos indekso prognozės kokybė

Įėjimo laipsnių agregacijos indekso s_{xy}^{CIN} prognozės kokybės tyrimo metu gauti rezultatai yra visai kitokie nei anksčiau nagrinėtų metodų atveju – ROC kreivės yra apverstos (žr. 18 pav.). Tai įprastai indikuoja faktą, kad prognozės metodas rezultatus skelbia priešingai t.y. tai, ką metodas skelbia kaip neegzistuojančią jungtį, iš tiesų jis turėtų skelbti priešingai – kaip egzistuojančią. Tai dažnai nėra blogas rezultatas, kadangi pakeitus taisyklę, pagal kurią yra surikiuojamos potencialios jungtys, galima gauti gerą kokybę atspindinčias ROC kreives (žr. 20a pav.).



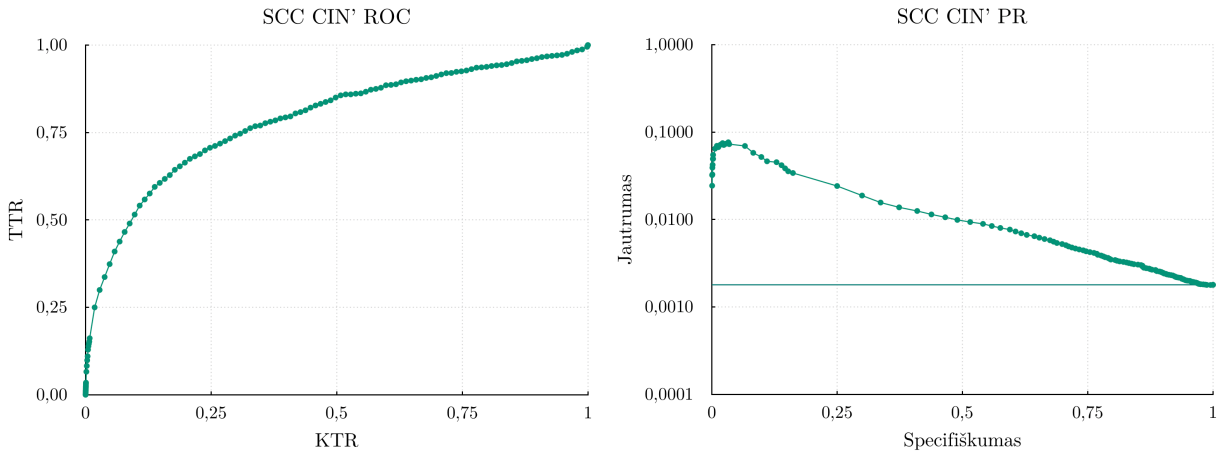
(a) (b)

18 pav. Tiriant įėjimo laipsnių agregacijos indekso s_{xy}^{CIN} prognozės kokybę gautos apverstos ROC kreivės. Pakeitus potencialių jungčių rikiavimo kryptį, galima gauti gerą prognozės kokybę indikuojančias kreives (žr. 20a pav.). ROC kreivės visiems tiriamiems tinklams pateikiamos priede Nr. 5.



(a) (b)

19 pav. Jautrumo-specifiškumo kreivės, gautos tiriant įėjimo laipsnių agregacijos indekso s_{xy}^{CIN} prognozės kokybę. Kreivės indikuoja labai prastus rezultatus: yra žemiau nei atsitiktinę prognozės kokybę žyminčios tiesės. Jautrumo-specifiškumo kreivės visiems tiriamiems tinklams pateikiamos priede Nr. 5.



(a) (b)

20 pav. ROC (a) ir jautrumo-specifiškumo (b) kreivės, gautos pakeitus potencialių jungčių rikiavimo kryptį. ROC kreivė indikuoja kokybišką prognozę, o jautrumo-specifiškumo kreivė rodo žymiai geresnius rezultatus nei originaliu s_{xy}^{CIN} indekso atveju (žr. 19a pav.).

Nors siekiant gerų jungčių prognozės ir bendrai dvejetainio klasifikavimo uždavinio rezultatų dažnai nėra klaidinga pakeisti tvarką, pagal kurią priskiriamos klasės, mūsų atveju tai nėra tinkamas sprendimas. Taip yra todėl, kad mums rūpi ne tiek kokybiška prognozė, kiek konkrečių prognozės metodų efektyvumas prognozuojant jungtis. Taigi, nors ir po rikiavimo tvarkos pakeitimo gavome ganėtinai kokybiškos prognozės rezultatus, šie rezultatai mums netinka tyrimo kontekste, nes metodas, pagal kurį šią prognozę atlikome nebėra įėjimo laipsnių agregacijos indeksas. Kaip matėme ankstesniuose darbo skyriuose, koeficientas $C^{IN}(D,k)$, kuriuo yra paremtas įėjimo laipsnių agregacijos indeksas, mažėja augant parametrui k, todėl potencialių jungčių rikiavimo pagal tikėtinumo indeksą tvarkos pakeitimas, reikštų šios mažėjančios tendencijos nepaisymą (tarsi teigtume, kad koeficientas $C^{IN}(D,k)$ didėja augant parametrui k, kas yra netiesa).

Čia galima iškelti keletą hipotezių, kalbančių apie įėjimo laipsnių agregacijos indekso s_{xy}^{CIN} kokybės tyrimo rezultatų kilmę:

1. Koeficiento $C^{IN}(D,k)$ mažėjimo tendencija augant parametrui nėra k patikima. Kaip matyti iš keleto diagramų (žr. 8a pav.), esant mažiems k įverčiams, koeficiento reikšmės nerodo labai aiškios mažėjimo tendencijos. Ir būtent šios jungtys gali atsirasti didėjimo tvarka surikiuoto sąrašo priekyje.

2. Realiuose tinkluose dažniausiai vyrauja laipsninis viršūnių laipsnių pasiskirstymas, tai reiškia, kad juose yra labai daug mažą laipsnį turinčių viršūnių ir labai nedaug – didelį. Surikiavus potencialias jungtis pagal požymį, susijusį su viršūnių laipsniais, tokioje masėje panašių ar vienodų įverčių gali būti prarandama prognozei reikšminga informacija.

3. Indeksas s_{xy}^{CIN} remiasi tinklo savybėmis, neturinčiomis reikšmės jungčių prognozei arba neatspindinčiomis prognozei reikšmingų tinklo charakteristikų.

5.4.4. Kokybės tyrimo rezultatų apibendrinimas

Tyrimo metu pastebėti ganėtinai geri kokybės įverčiai jungčių prognozei taikant liudininkų s_{xy}^{CW} ir bendrų interesų s_{xy}^{CT} indeksus (žr. 11 ir 12 lenteles). Abiems indeksams ROC kreivės yra ryškiai nutolusios nuo atsitiktinei prognozei prognozei būdingų rezultatų, o AUROC metrika atitinkamai įgyja gerą prognozės kokybę indikuojančias reikšmes.

Liudininkų bei bendrų interesų indeksams ganėtinai kokybiškus rezultatus rodo ir jautrumo-specifiškumo kreivės. Nors jautrumas sparčiai mažėja augant specifiškumui, pastarajam esant mažam (mažas kiekis potencialių jungčių prognozuojamos kaip egzistuojančios), matomi ypač geri jautrumo įverčiai. Kadangi dauguma realaus pasaulio tinklų yra reti, tokie rezultatai atrodo prasmingi: tik mažas kiekis visų potencialių jungčių iš tiesų ir susiformuos. Atitinkamai ir AUPR įverčiai indikuoja už atsitiktinę prognozę daug geresnius rezultatus.

Įėjimo laipsnių agregacijos indekso s_{xy}^{CIN} prognozės kokybės tyrimo rezultatai, tuo tarpu, atskleidė visai kitokius rezultatus – metodas jungtis prognozuoja atvirkščiai t.y. tos jungtys, kurios metodo yra skelbiamos kaip egzistuojančios, iš tiesų neegzistuoja ir priešingai. Taigi, šiuo atveju jungčių prognozavimas nagrinėjant mažus įėjimo laipsnius turinčias viršūnes (kurių retuose tinkluose yra labai daug) nepadeda kokybiškai prognozuoti tinklo jungčių.

11 lentelė. AUROC metrikos tirtuose tinkluose kiekvienam iš nagrinėjamų jungčių tikėtinumo indeksų. Didžiausi įverčiai tinklui paryškinti juodžiau.

Indeksas	AUROC$_{WIK}$	AUROC$_{SCC}$	AUROC$_{SPA}$	AUROC$_{SPC}$	AUROC$_{SSA}$	AUROC$_{SSC}$
s_{xy}^{CW}	**0,91**	**0,86**	0,76	0,84	0,74	0,79
s_{xy}^{CT}	**0,91**	0,85	**0,85**	**0,91**	**0,84**	**0,86**
s_{xy}^{CIN}	0,08	0,22	0,18	0,14	0,24	0,2

12 lentelė. AUPR metrikos tirtuose tinkluose kiekvienam iš nagrinėjamų jungčių tikėtinumo indeksų. s_{xy}^{RAND} žymi atsitiktinai jungtis prognozuojantį metodą. Didžiausi įverčiai tinklui paryškinti juodžiau.

Indeksas	AUPR$_{WIK}$	AUPR$_{SCC}$	AUPR$_{SPA}$	AUPR$_{SPC}$	AUPR$_{SSA}$	AUPR$_{SSC}$
s_{xy}^{CW}	0,013	**0,032**	**0,01**	**0,018**	0,005	0,018
s_{xy}^{CT}	**0,019**	0,023	0,004	0,013	**0,008**	**0,027**
s_{xy}^{CIN}	$5{,}2 \cdot 10^{-4}$	10^{-3}	$1{,}3 \cdot 10^{-4}$	$6{,}5 \cdot 10^{-5}$	$2{,}5 \cdot 10^{-4}$	$1{,}9 \cdot 10^{-4}$
s_{xy}^{RAND}	$9{,}2 \cdot 10^{-4}$	$1{,}8 \cdot 10^{-3}$	$2{,}4 \cdot 10^{-4}$	10^{-4}	$4 \cdot 10^{-4}$	$3 \cdot 10^{-4}$

Rezultatai ir išvados

Rezultatai

Darbe apibrėžtas ir išanalizuotas jungčių prognozės uždavinys bei jo tipai. Taip pat pristatytas šiems uždaviniams spręsti taikytinų metodų vertinimo karkasas bei aibė šiam vertinimui tinkamų metrikų. Čia pastebėti ROC ir jautrumo-specifiškumo kreivių bei plotų po jais (kaip vertinimo metrikų) pranašumai.

Vėliau, remiantis literatūra, išnagrinėtas rinkinys dažniau sutinkamų ir praktikoje taikomų jungčių prognozės metodų, kurie darbe suskirstyti į dvi esmines grupes: lokalius ir globalius pagal tai, su kokio lygio informacija tinkle jie dirba. Be to, pateikti ir išanalizuoti empiriniai dalies nagrinėjamų prognozės metodų kokybės įverčiai realiuose tinkluose. Šie įverčiai palyginti tarpusavyje ir pastebėtas globalių metodų pranašumas.

Galiausiai, išnagrinėta jungčių prognozės orientuotuose tinkluose specifika bei skirtumai nuo neorientuotų tinklų atvejo. Taip pat išskirti esminiai jungčių prognozės orientuotuose tinkluose iššūkiai.

Antrojoje ir trečiojoje darbo dalyse išanalizuota tiriama orientuoto tinklo klasterizacijos samprata su jos praplėtimais bei **apibrėžti** trys **nauji** šia samprata paremti jungčių prognozės metodai: liudininkų, bendrų interesų ir įėjimo laipsnių agregacijos jungčių **tikėtinumo indeksai**.

Ketvirtojoje dalyje aprašyti tyrimui surinkti duomenys. Tarp šių duomenų pateikiama pora populiarių akademinėje bendruomenėje tinklų bei keturi iš specialiai šiam darbui surinktų duomenų **sumodeliuoti tinklai**. Pastarieji tinklai šioje darbo dalyje analizuojami detaliau: nagrinėjama jų dvidalė prigimtis, viršūnių laipsnių skirstiniai, jungiųjų komponenčių analizė.

Paskutinėje (penktojoje) šio darbo dalyje atliktas jungčių prognozės **kokybės**, paremtos apibrėžtais jungčių tikėtinumo indeksais, **tyrimas**. Taip pat aprašyti šiam tyrimui paruošti įrankiai, įgyvendinti algoritmai, pateikta išsami minėtųjų jungčių tikėtinumo **indeksų** priklausomybės nuo jų parametrų **analizė**. Šioje darbo dalyje taip pat išsamiai išdėstyta atlikto tyrimo eiga ir metodika, pateikti detalūs kiekvieno prognozės metodo tyrimo rezultatai su jų vertinimu.

Išvados

Darbe apibrėžtus jungčių tikėtinumo indeksus empiriškai ištyrus visuose šešiuose nagrinėjamuose tinkluose pastebėti aiškias tendencijas rodantys rezultatai, pagal kuriuos galima padaryti **išvadas** apie šių indeksų jungčių prognozės kokybę. Eksperimentų metu pastebėti **aukšti** kokybės metrikų **įverčiai**, kada jungčių prognozei yra taikomi liudininkų ir bendrų interesų indeksai. Nors jautrumo-specifiškumo kreivėse jautrumas sparčiai mažėja augant specifiškumui, pastarajam esant mažam (mažas kiekis potencialių jungčių prognozuojamos kaip egzistuojančios), matomi

ypač geri jautrumo įverčiai. Kadangi dauguma realaus pasaulio tinklų yra reti, tokie rezultatai atrodo prasmingi: tik mažas kiekis visų potencialių jungčių iš tiesų ir susiformuos. Atitinkamai ir AUROC bei AUPR įverčiai indikuoja už atsitiktinę prognozę daug geresnius rezultatus.

Įėjimo laipsnių agregacijos indeksas, tuo tarpu, atskleidė visai kitokius rezultatus – šis metodas jungtis prognozuoja atvirkščiai t.y. jungtys, kurias metodas skelbia egzistuojančiomis, iš tiesų turėtų būti skelbiamos neegzistuojančiomis ir priešingai. Taigi, šiuo atveju jungčių prognozavimas nagrinėjant mažus įėjimo laipsnius turinčias viršūnes (kurių retuose tinkluose yra labai daug) nepadeda kokybiškai prognozuoti tinklo jungčių. Tokių rezultatų prigimtis nėra iki galo aiški, tačiau gali būti susijusi su faktu, kad mažo laipsnio viršūnių realiuose tinkluose yra labai daug ir šioje daugybėje prarandama svarbi topologinė informacija. Viena iš priežasčių gali būti ir klasterizacijos koeficiento, kuriuo remiasi įėjimo laipsnių agregacijos indeksas, priklausomybės nuo jo parametro (viršūnės įėjimo laipsnio) nestabilumas esant mažoms šio parametro reikšmėms. Taip pat negalima atmesti ir galimybės, kad įėjimo laipsnių agregacijos indeksas remiasi tinklo savybėmis, neturinčiomis reikšmės jungčių prognozei arba neatspindinčiomis prognozei reikšmingų tinklo charakteristikų. Įėjimo laipsnių agregacijos indekso detalesnė analizė ir tobulinimas gali būti viena iš tolimesnio darbo krypčių.

Literatūra

[AA03] L. A. Adamic ir E. Adar. Friends and neighbors on the web. *Social networks*:211–230, 3, 2003-07.

[AB02] R. Albert ir A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 1, 2002-01.

[BL16] M. Bloznelis ir L. Leskelä. Diclique clustering in a directed random graph. *In A. Bonato, F. Chung Graham, P. Pralat (Eds.): Algorithms and Models for the Web Graph - 13th International Workshop, WAW 2016, Lecture Notes in Computer Science 10088, Springer*:22–33, 2016.

[Dav17] T. Davidovič. *Realių tinklų kaimynystės ryšių statistinė priklausomybė: empiriniai tyrimai ir modeliavimas*. Bakalauro baigiamasis darbas, vad. M. Bloznelis, Vilniaus universitetas, 2017.

[FV13] D. M. Fragkiskos ir M. Vazirgiannis. Clustering and community detection in directed networks: a survey. *CoRR*, abs/1308.0971, 2013.

[GMC+15] F. Gao, K. Musial, C. Cooper ir S. Tsoka. Link prediction methods and their accuracy for different social networks and network metrics, 2015-11.

[HCS+06] M. Al Hasan, V. Chaoji, S. Salem ir M. Zaki. Link prediction using supervised learning. *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.

[Hua06] Z. Huang. Link prediction based on graph topology: the predictive value of the generalized clustering coefficient, 2006.

[Jac01] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 1901.

[Kat53] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953-03.

[LHK10] J. Leskovec, D. Huttenlocher ir J. Kleinberg. Signed networks in social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 1361–1370, New York, NY, USA, 2010.

[LK07] D. Liben-Nowell ir J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. ISSN: 1532-2890.

[LLC10] R. N. Lichtenwalter, J. T. Lussier ir N. V. Chawla. New perspectives and methods in link prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, p. 243–252, 2010.

[LW71] F. Lorrain ir H. C. White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 49, 1971.

[LZ11] L. Lü ir T. Zhou. Link prediction in complex networks: a survey. *Physica A: statistical mechanics and its applications*, 390:1150–1170, 6, 2011.

[New01] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64, 2, 2001-07.

[New03] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45:16, 2003.

[PJ14] J. Pengsheng ir J. Jiashun. Coauthorship and citation networks for statisticians. *Annals of Applied Statistics*, 10, 2014-10.

[Puj15] M. Pujari. *Link Prediction in Large-scale Complex Networks*. Disertacija, Université Paris Nord, 2015.

[SM86] G. Salton ir M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[Vai17] R. Vainickas. *Realių tinklų kaimynystės ryšių statistinė priklausomybė: empiriniai tyrimai ir modeliavimas*. Bakalauro baigiamasis darbas, vad. M. Bloznelis, Vilniaus universitetas, 2017.

[WLZ17] Z. Wu, Y. Lin ir Y. Zhao. Improving local clustering based top-l link prediction methods via asymmetric link clustering information. *Physica A: Statistical Mechanics and its Applications*:1–14, 2017-11.

[WMR+15] Z. Wu, G. Menichetti, C. Rahmede ir G. Bianconi. Emergent complex network geometry. *Scientific Reports*, 5:10073, 2015-10.

[WS98] J. D. Watts ir S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[ZYW14] J. Zhang, P. S. Yu ir O. E. Wolfson. Link prediction across heterogeneous social networks: a survey. 2014.

[ZLW+13] Q.-M. Zhang, L. Lü, W.-Q. W.-Q. Wang, Yu-Xiao ir T. Zhou. Potential theory for directed networks. *PLOS ONE*, 8(2):1–8, 2013-02.

[ZLZ09] T. Zhou, L. Lü ir Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71:623–630, 4, 2009-10.

[ZRM+09] T. Zhou, J. Ren, M. Medo ir Y.-C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76:046115, 4, 2009-10.

Priedas Nr. 1

Tyrimo medžiaga ir programinė įranga

Visi tyrimo eksperimentai atlikti programine įranga, sukurta *Go*[14] programavimo aplinkoje. Naudota *Go* 1.12.2 versija *darwin/amd64* architektūroje. Darbui papildomai naudota ir mokslinė *Gonum*[15] biblioteka.

Darbo metu surinkti duomenys, sumodeliuoti tinklai, visa sukurta programinė įranga bei detalūs kiekvieno tyrimo rezultatai pasiekiami *Google* debesyje[16]. Debesyje pateikiamo priedo struktūra:

1. `gonum.org.zip` – archyvas su *Gonum* biblioteka. Šią biblioteką reikėtų turėti `GOPATH` kataloge.
2. `master_thesis_material.zip` – archyvas su darbo metu surinktais duomenimis, sumodeliuotais tinklais, visa sukurta programine įranga bei detaliais kiekvieno tyrimo rezultatais.
    - `README.md` – darbo metu parašytų pagalbinių bibliotekų aprašas su naudojimo pavyzdžiais.
    - `lib` – katalogas su darbo metu parašytų pagalbinių bibliotekų išeities kodu.
    - `programs` - katalogas su eksperimentų metu parašytų programų, tokių kaip tinklo metrikų ir koeficientų skaičiavimas, jungčių prognozavimas ar rezultatų vertinimas išeities kodu.
    - `config` – katalogas su eksperimentų programų konfigūracijų dokumentais.
    - `data` – katalogas su surinktais duomenimis, sumodeliuotais tinklais, jų metrikomis bei detaliais tyrimų rezultatais.
        - `physics/bipartite` – fizikos *StackExchange* tinklai, jų metrikos ir tyrimo rezultatai.
        - `stats/bipartite` – statistikos *StackExchange* tinklai, jų metrikos ir tyrimo rezultatai.
        - `scc/bipartite` – statistikų citavimo tinklas, jo metrikos ir tyrimo rezultatai.
        - `wiki_vote` – *Wikipedia* balsavimo tinklas, jo metrikos ir tyrimo rezultatai.

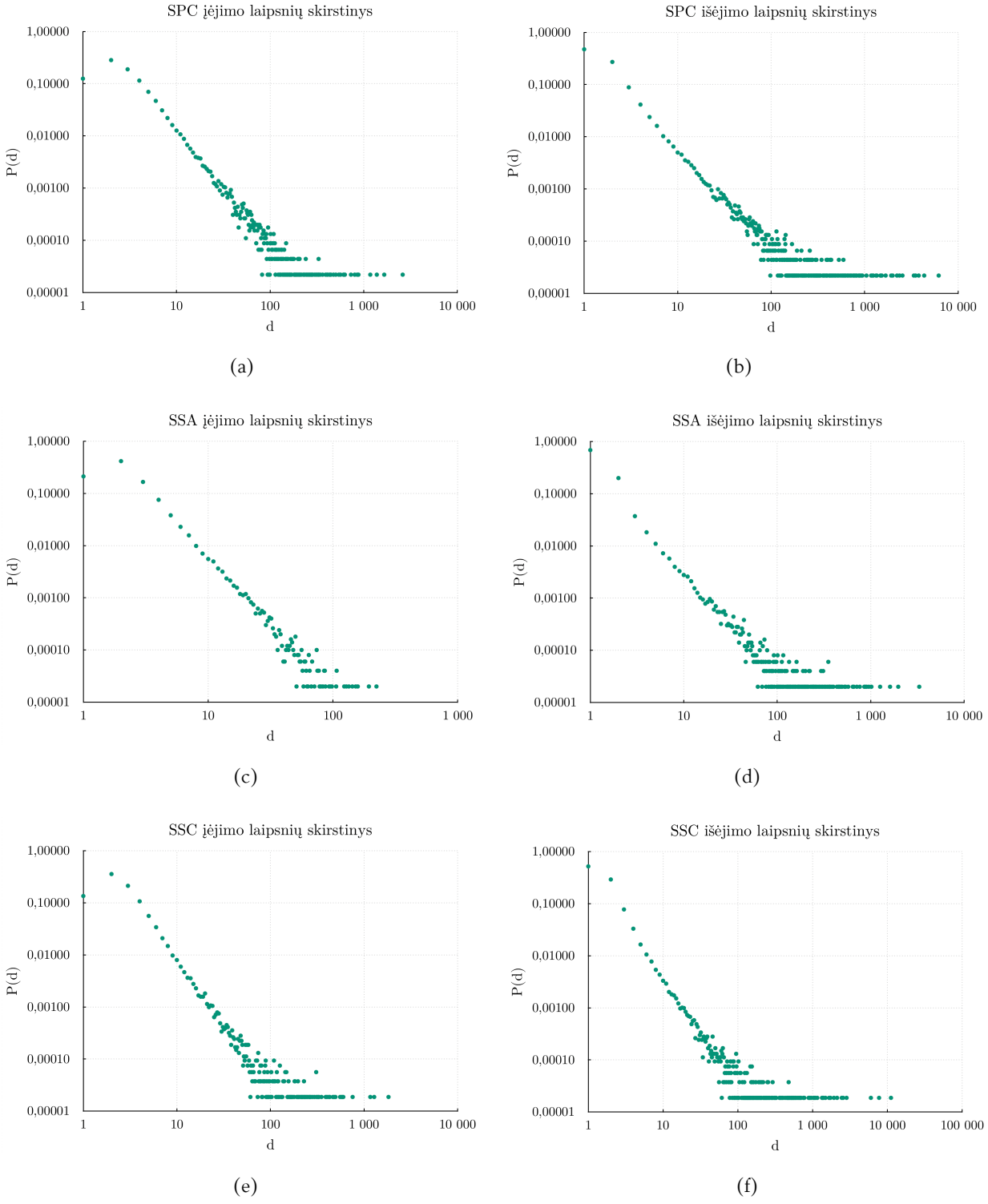
Katalogų `physics/bipartite`, `stats/bipartite`, `scc/bipartite`, `wiki_vote` struktūra:

- `readme.md` – tinklų aprašai.
- `report.md` – tinklų metrikos.
- `graph` – katalogas su dvidalio tinklo dalių ir iš jų išvestinių tinklų briaunų sąrašais.
- `degree_frequencies` – katalogas su tinklų viršūnių laipsnių dažnių lentelėmis.
- `prognosis/samples` - katalogas su iš tiriamų tinklų parengtais mokymo ir testavimo potinkliais.
- `coefficients/full_graph` – katalogas su tiriamų koeficientų įverčiais nesumažintuose tinkluose.
- `coefficients/sample` – katalogas su tiriamų koeficientų įverčiais mokymo potinkliuose.
- `prognosis/validation` - katalogas su prognozės kokybės vertinimo rezultatais.

[14] `https://golang.org/`
[15] `https://www.gonum.org/`
[16] `https://drive.google.com/drive/folders/1tSbrtXOydOHIJYVMKWSIVBJylei_etTd?usp=sharing`

Priedas Nr. 2

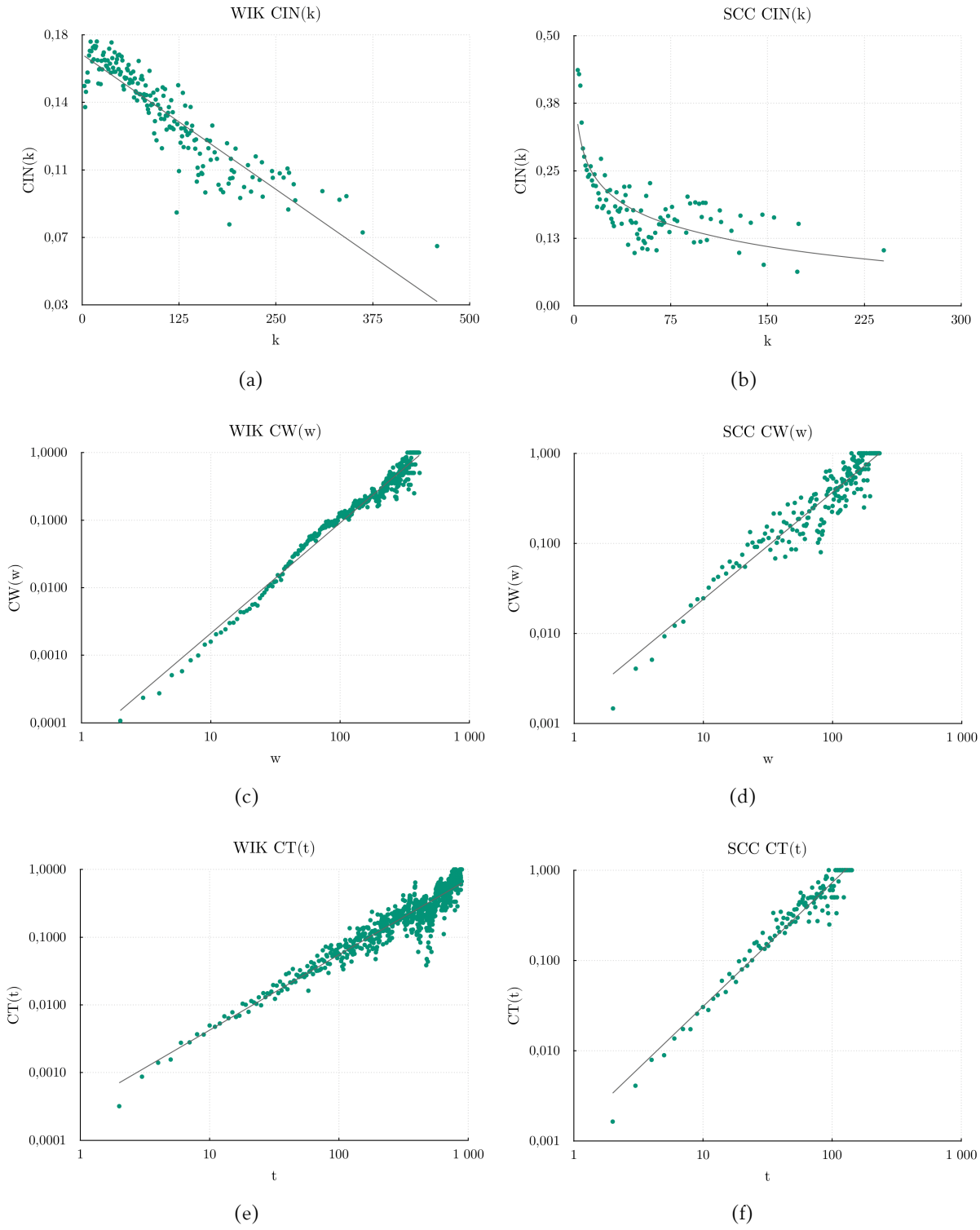
Viršūnių laipsnių pasiskirstymas modeliuotuose tinkluose



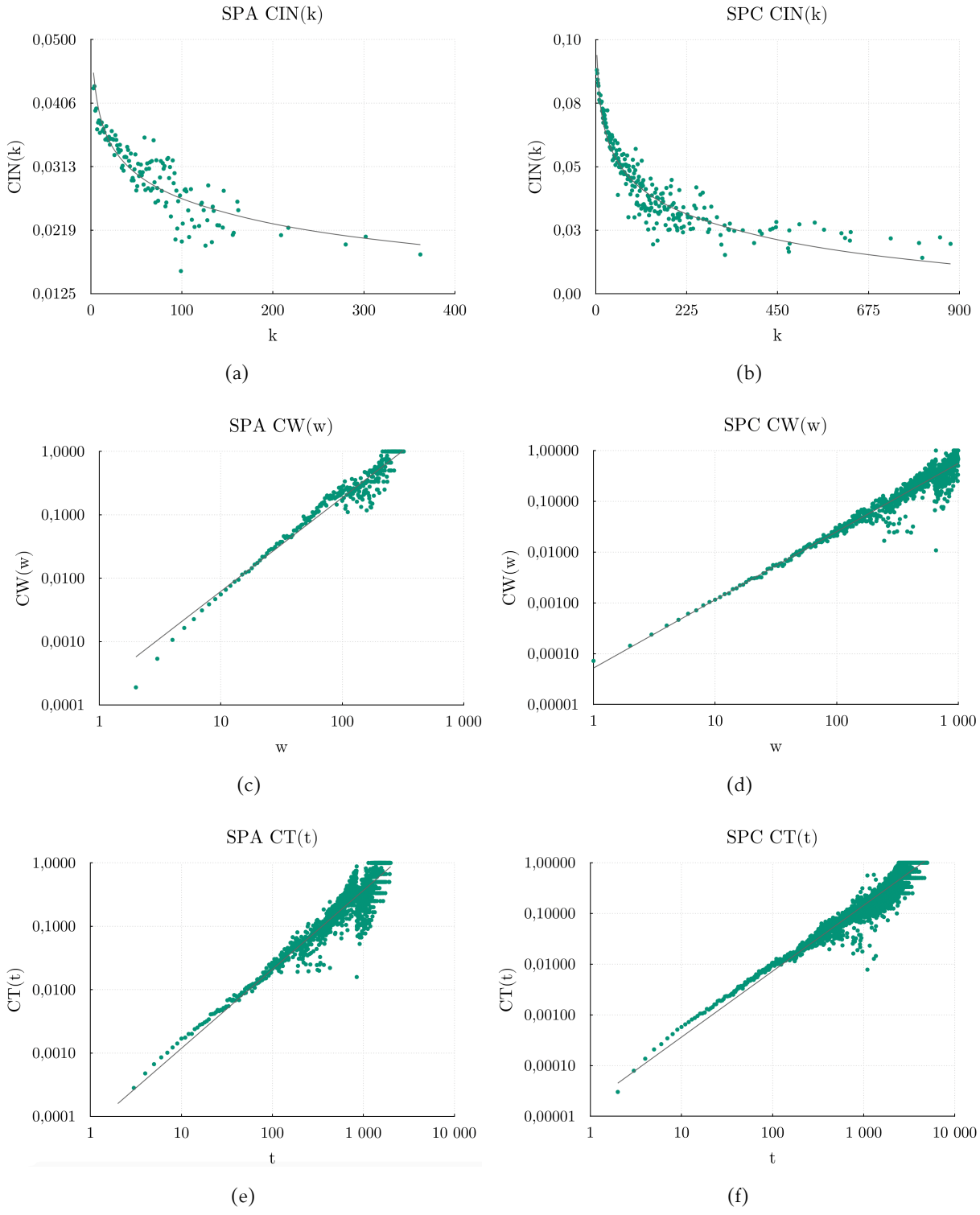
21 pav. Viršūnių laipsnių skirstiniai tyrimui sumodeliuotuose tinkluose. Duomenis taškinėse diagramose konvertavus į logaritmo pagrindu dešimt skalę matoma į tiesę panaši priklausomybė. Tokia tendencija yra būdinga laipsniniam skirstiniui.

Priedas Nr. 3

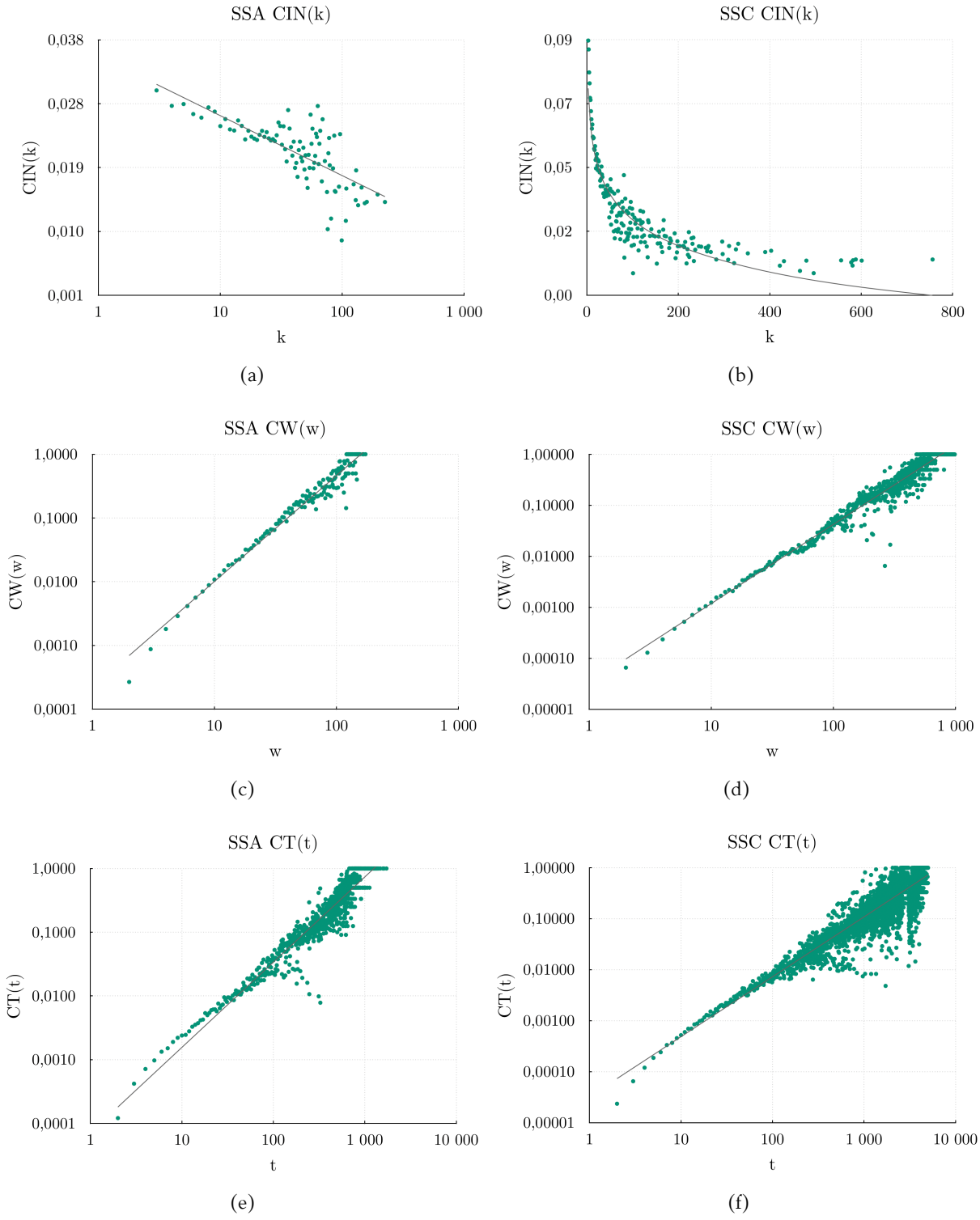
Koeficientų priklausomybė nuo jų parametrų tiriamuose tinkluose



22 pav. Koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ kitimo tendencijų priklausomybė nuo jų parametrų tiriamuose (nesumažintuose) tinkluose *WIK* ir *SCC*. Siekiant geriau atskleisti koeficientų priklausomybės tendencijas, taškinių diagramų (c), (d), (e) ir (f) abi ašys paverstos į logaritmo pagrindu dešimt skalę.



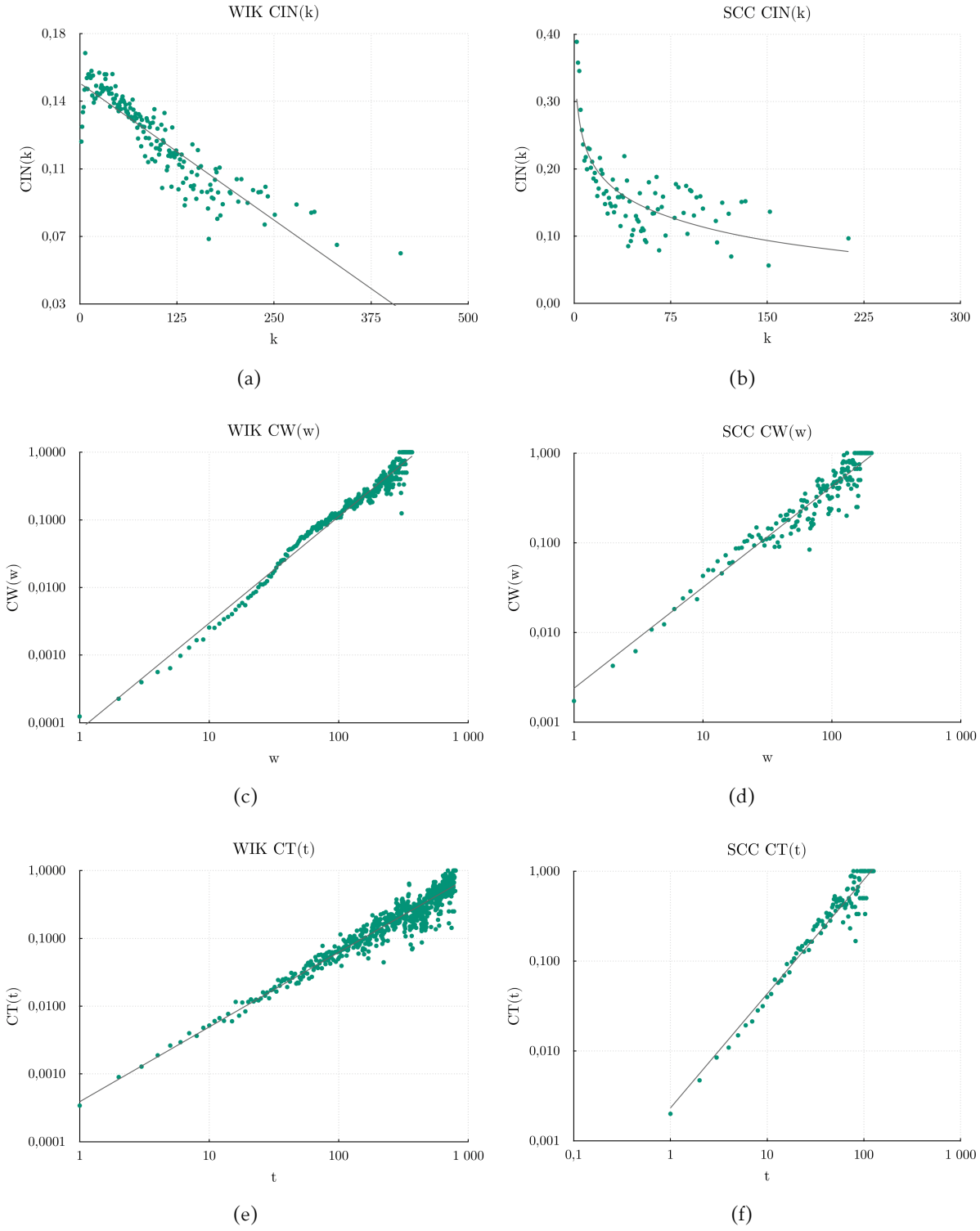
23 pav. Koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ kitimo tendencijų priklausomybė nuo jų parametrų tiriamuose (nesumažintuose) tinkluose *SPA* ir *SPC*. Siekiant geriau atskleisti koeficientų priklausomybės tendencijas, taškinių diagramų (c), (d), (e) ir (f) abi ašys paverstos į logaritmo pagrindu dešimt skalę.



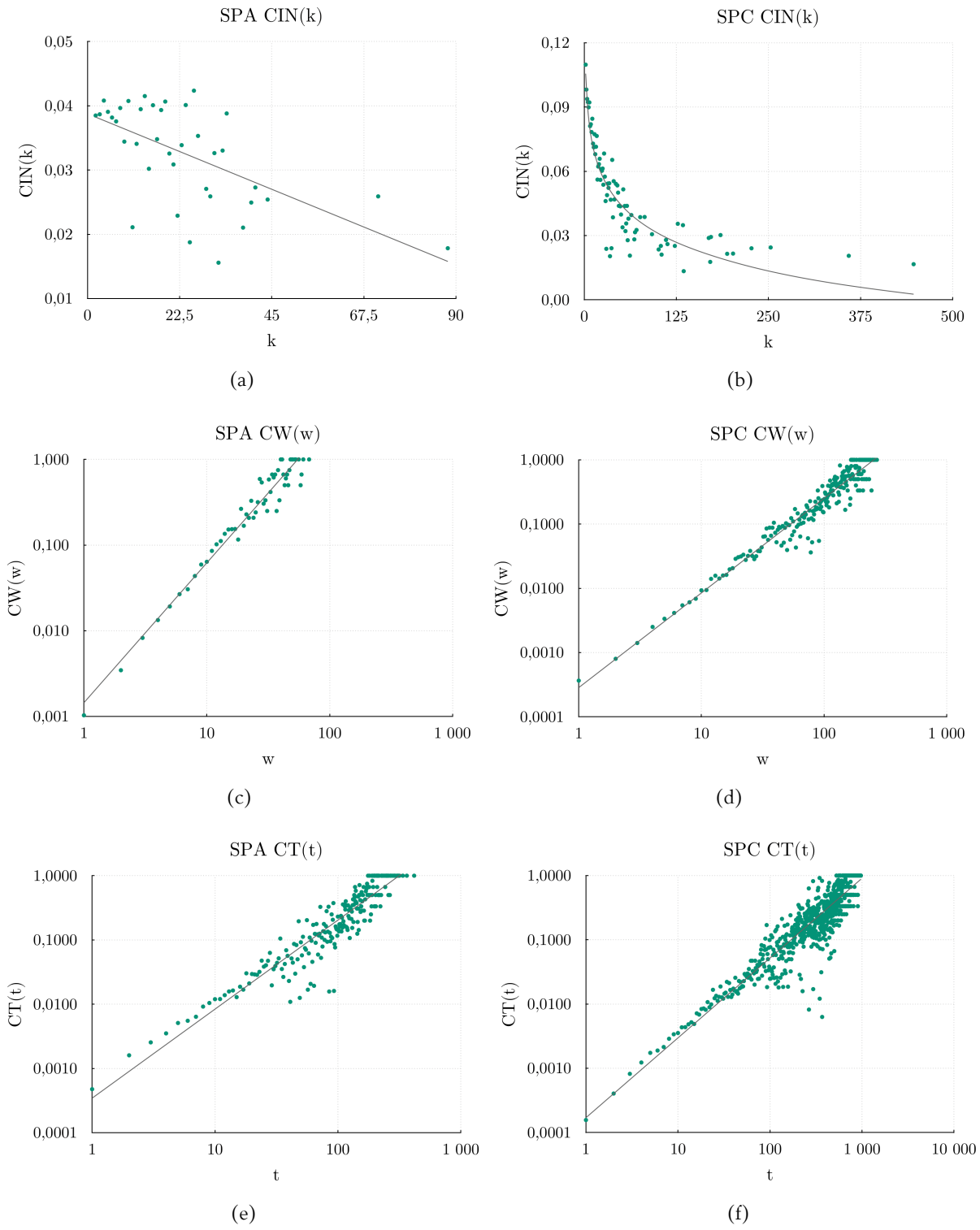
24 pav. Koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ kitimo tendencijų priklausomybė nuo jų parametrų tiriamuose (nesumažintuose) tinkluose *SSA* ir *SSC*. Siekiant geriau atskleisti koeficientų priklausomybės tendencijas, taškinių diagramų (c), (d), (e) ir (f) abi ašys, o diagramos (a) k ašis paverstos į logaritmo pagrindu dešimt skalę.

Priedas Nr. 4

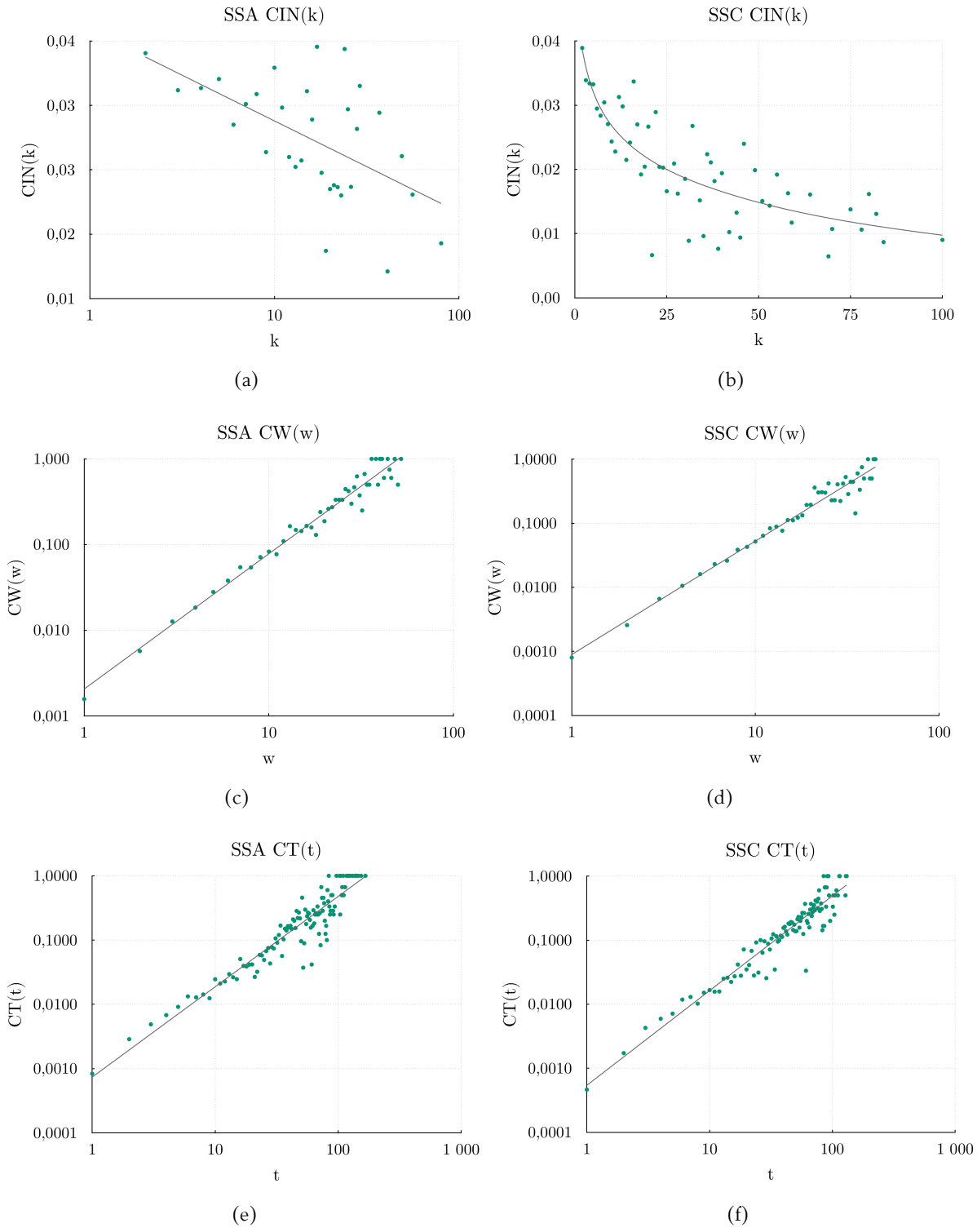
Koeficientų priklausomybė nuo jų parametrų mokymo potinkliuose



25 pav. Koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ kitimo tendencijų priklausomybė nuo jų parametrų mokymo potinkliuose, gautuose sumažinus tinklus *WIK* ir *SCC*. Siekiant geriau atskleisti koeficientų priklausomybės tendencijas, taškinių diagramų (c), (d), (e) ir (f) abi ašys paverstos į logaritmo pagrindu dešimt skalę.



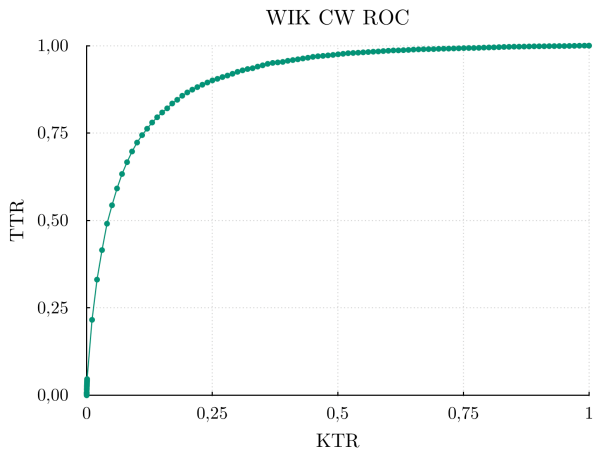
26 pav. Koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ kitimo tendencijų priklausomybė nuo jų parametrų mokymo potinkliuose, gautuose sumažinus tinklus *SPA* ir *SPC*. Siekiant geriau atskleisti koeficientų priklausomybės tendencijas, taškinių diagramų (c), (d), (e) ir (f) abi ašys paverstos į logaritmo pagrindu dešimt skalę.



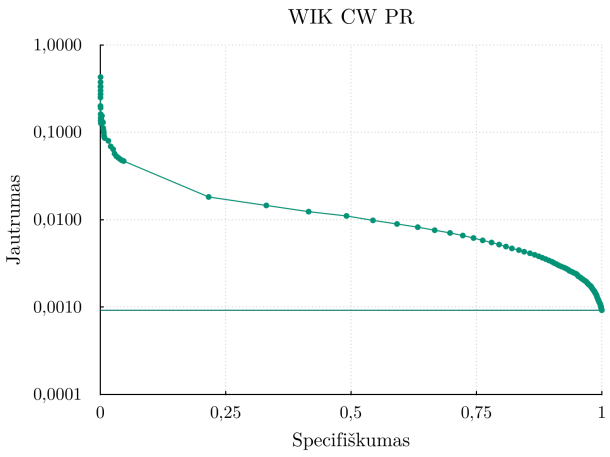
27 pav. Koeficientų $C^{IN}(D,k)$, $C^{W}(D,w)$ ir $C^{T}(D,t)$ kitimo tendencijų priklausomybė nuo jų parametrų mokymo potinkliuose, gautuose sumažinus tinklus *SSA* ir *SSC*. Siekiant geriau atskleisti koeficientų priklausomybės tendencijas, taškinių diagramų (c), (d), (e) ir (f) abi ašys bei diagramos (a) k ašis paverstos į logaritmo pagrindu dešimt skalę

Priedas Nr. 5

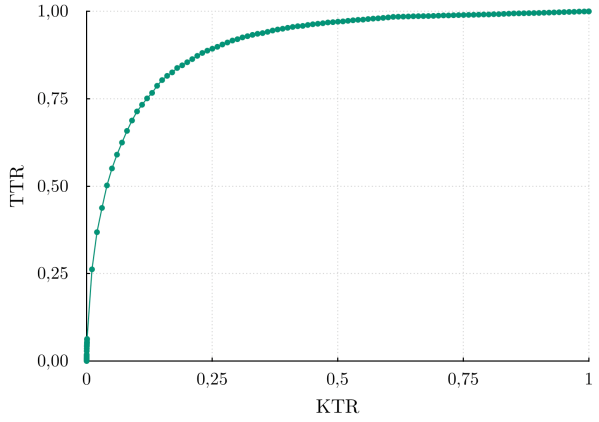
Prognozės kokybės kreivės tiriamiems jungčių tikėtinumo indeksams



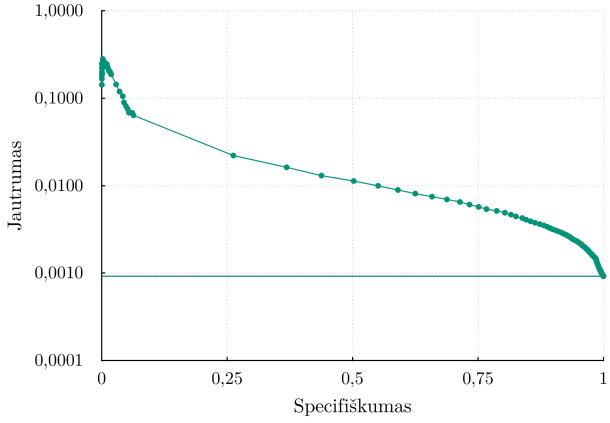
(a) s_{xy}^{CW}



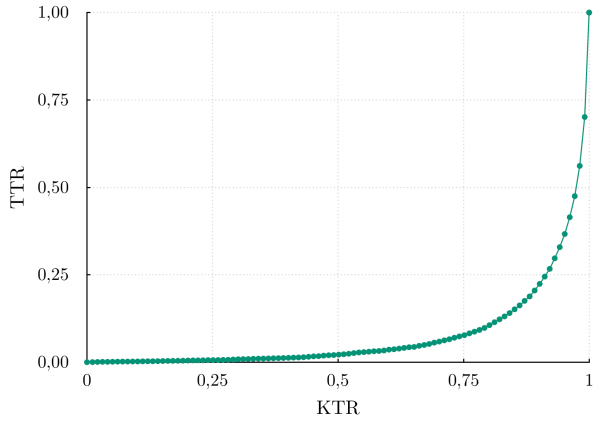
(b) s_{xy}^{CW}



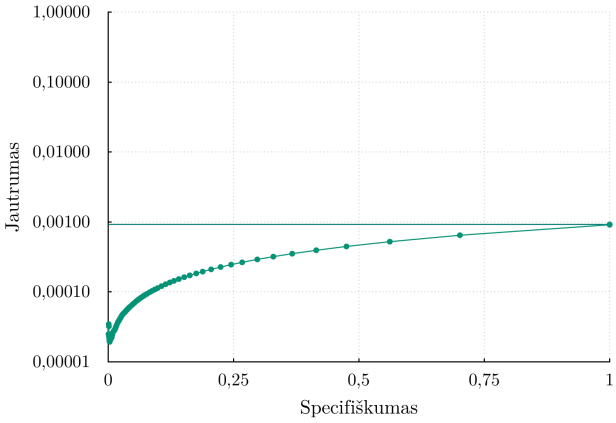
(c) s_{xy}^{CT}



(d) s_{xy}^{CT}

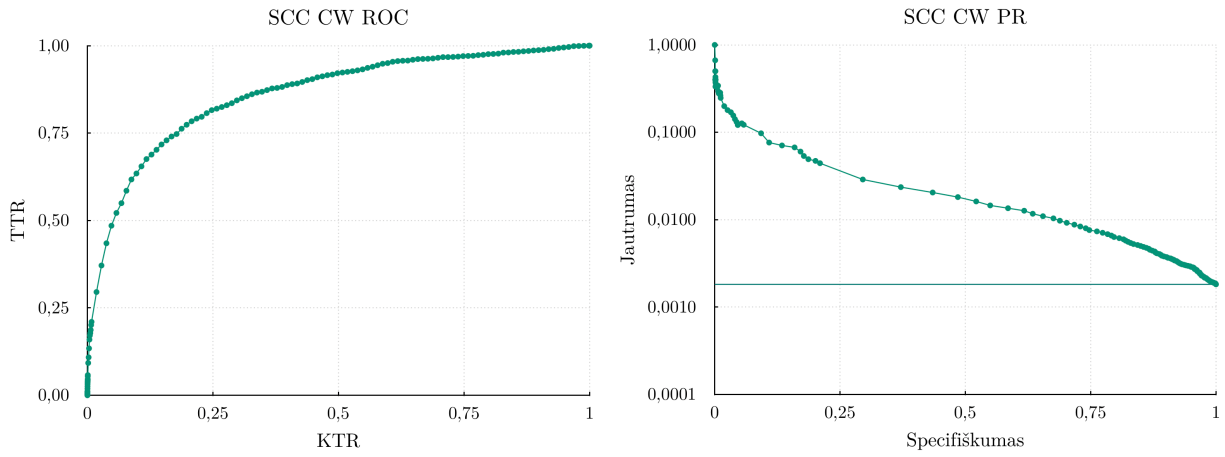


(e) s_{xy}^{CIN}



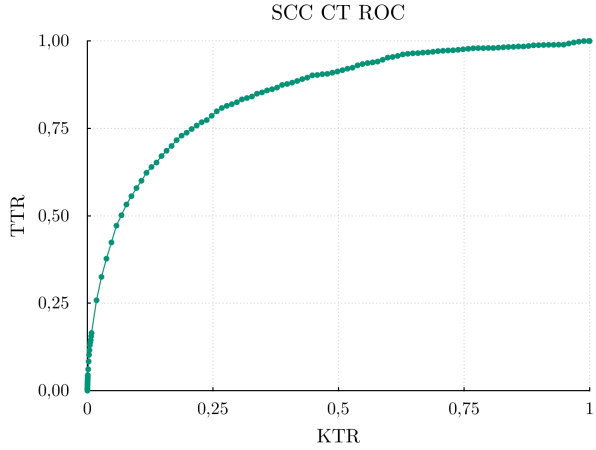
(f) s_{xy}^{CIN}

28 pav. Prognozės, naudojantis indeksais s_{xy}^{CW}, s_{xy}^{CT} ir s_{xy}^{CIN}, kokybės metrikos (ROC ir jautrumo-specifiškumo kreivės) mokymo ir testavimo potinkliams, paruoštiems iš tinklo *WIK*. Jautrumo-specifiškumo kreivių jautrumo ašys paverstos į logaritmo pagrindu dešimt skalę.

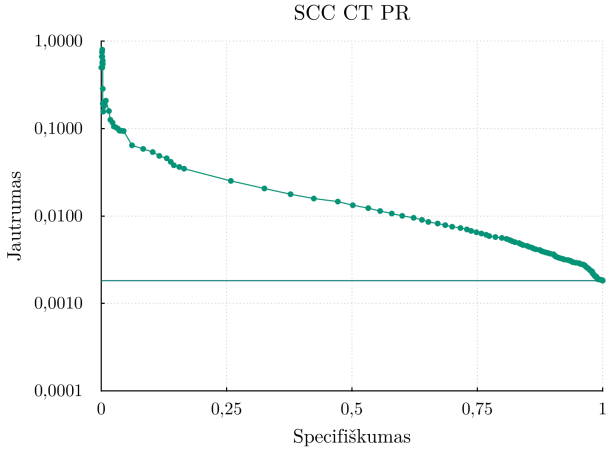


(a) s_{xy}^{CW}

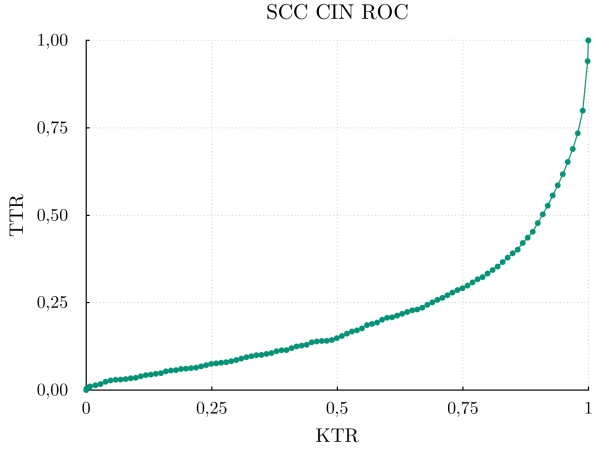
(b) s_{xy}^{CW}



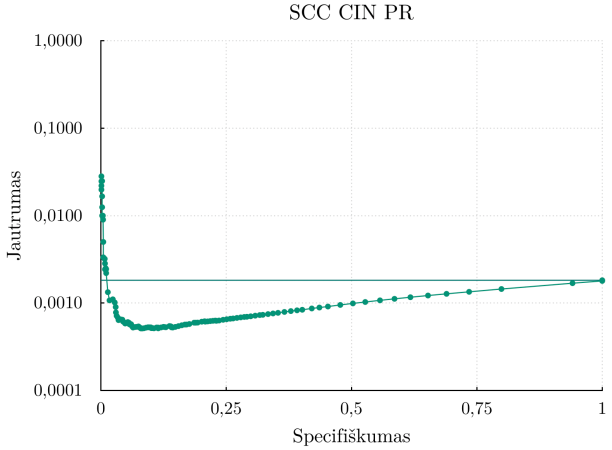
(c) s_{xy}^{CT}



(d) s_{xy}^{CT}

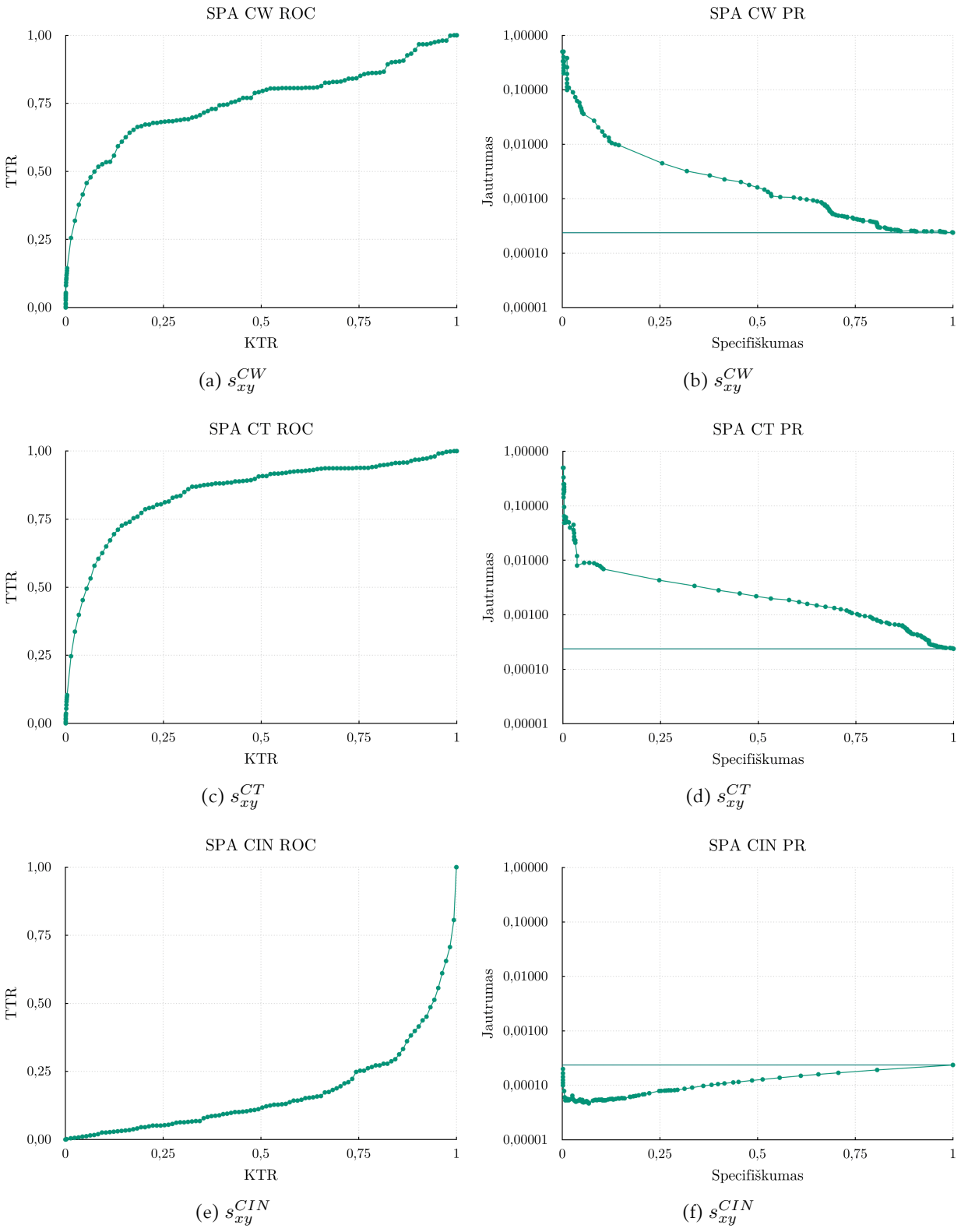


(e) s_{xy}^{CIN}



(f) s_{xy}^{CIN}

29 pav. Prognozės, naudojantis indeksais s_{xy}^{CW}, s_{xy}^{CT} ir s_{xy}^{CIN}, kokybės metrikos (ROC ir jautrumo-specifiškumo kreivės) mokymo ir testavimo potinkliams, paruoštiems iš tinklo *SCC*. Jautrumo-specifiškumo kreivių jautrumo ašys paverstos į logaritmo pagrindu dešimt skalę.

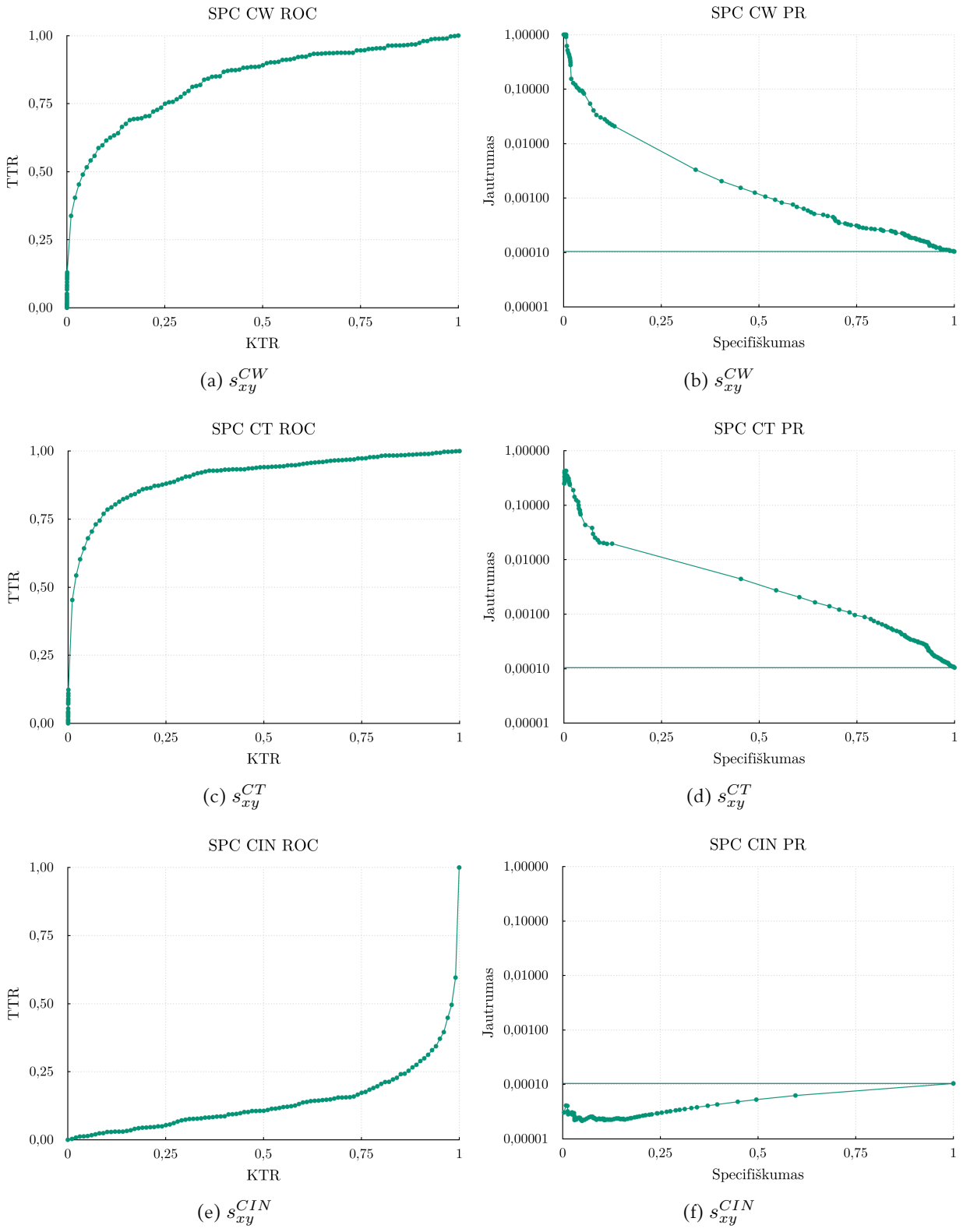


(a) s_{xy}^{CW} (b) s_{xy}^{CW}

(c) s_{xy}^{CT} (d) s_{xy}^{CT}

(e) s_{xy}^{CIN} (f) s_{xy}^{CIN}

30 pav. Prognozės, naudojantis indeksais s_{xy}^{CW}, s_{xy}^{CT} ir s_{xy}^{CIN}, kokybės metrikos (ROC ir jautrumo-specifiškumo kreivės) mokymo ir testavimo potinkliams, paruoštiems iš tinklo *SPA*. Jautrumo-specifiškumo kreivių jautrumo ašys paverstos į logaritmo pagrindu dešimt skalę.

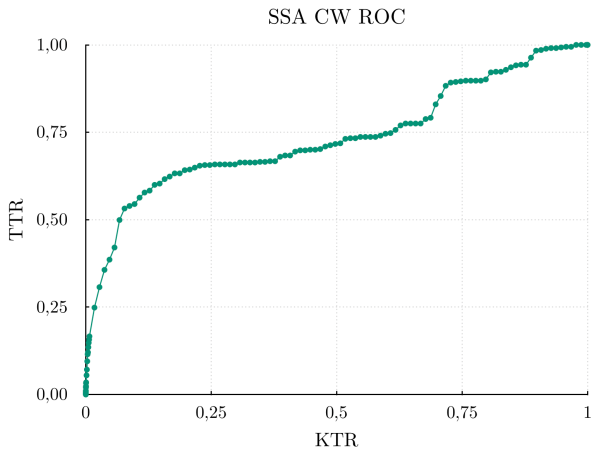


(a) s_{xy}^{CW} (b) s_{xy}^{CW}

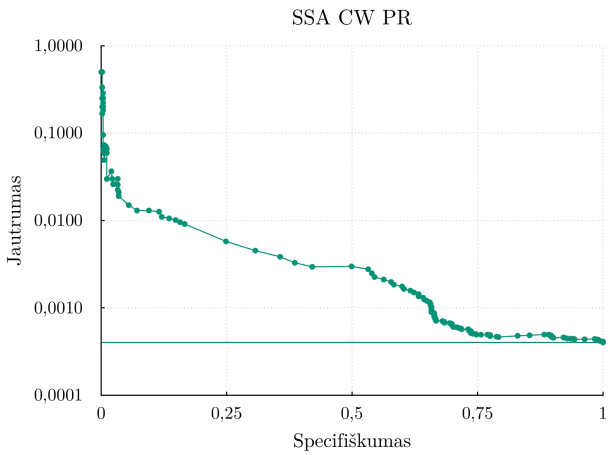
(c) s_{xy}^{CT} (d) s_{xy}^{CT}

(e) s_{xy}^{CIN} (f) s_{xy}^{CIN}

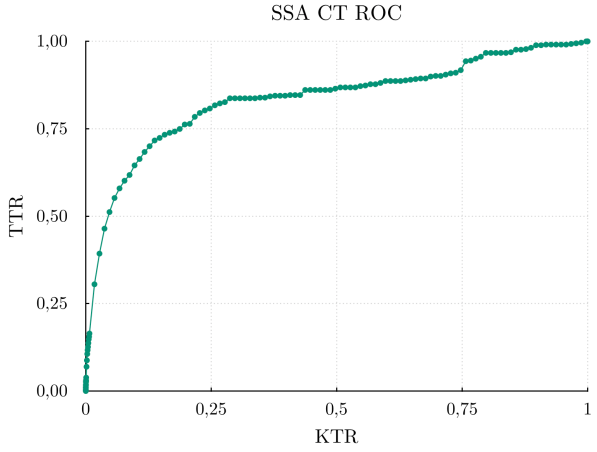
31 pav. Prognozės, naudojantis indeksais s_{xy}^{CW}, s_{xy}^{CT} ir s_{xy}^{CIN}, kokybės metrikos (ROC ir jautrumo-specifiškumo kreivės) mokymo ir testavimo potinkliams, paruoštiems iš tinklo *SPC*. Jautrumo-specifiškumo kreivių jautrumo ašys paverstos į logaritmo pagrindu dešimt skalę.



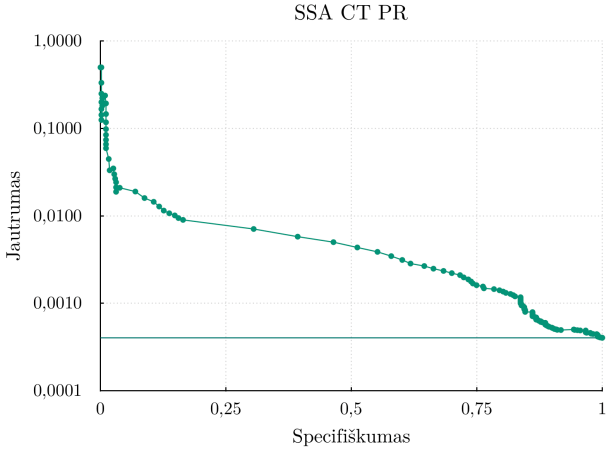
(a) s_{xy}^{CW}



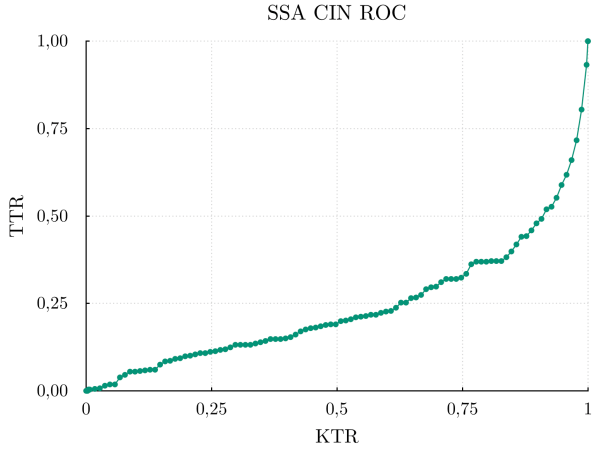
(b) s_{xy}^{CW}



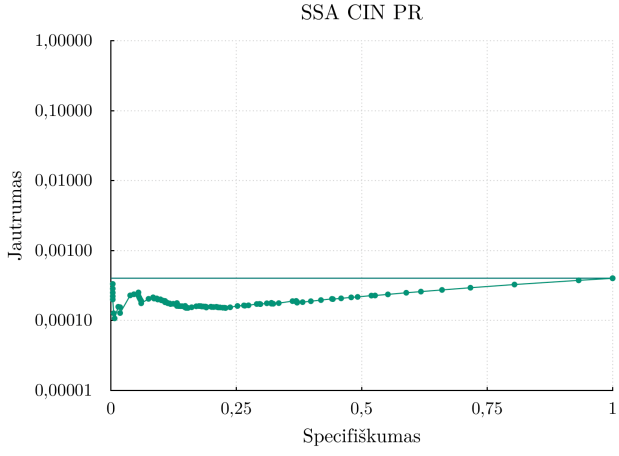
(c) s_{xy}^{CT}



(d) s_{xy}^{CT}

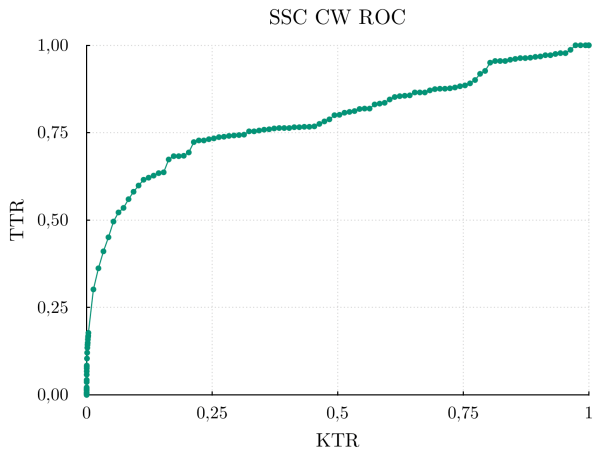


(e) s_{xy}^{CIN}

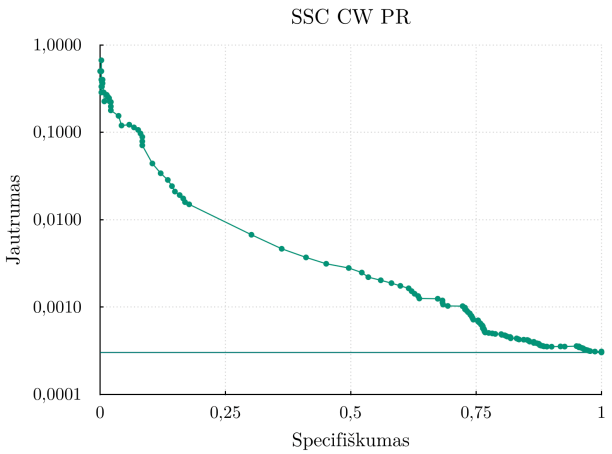


(f) s_{xy}^{CIN}

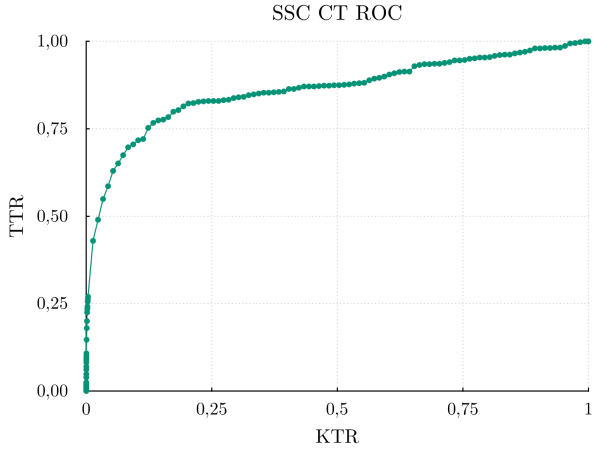
32 pav. Prognozės, naudojantis indeksais s_{xy}^{CW}, s_{xy}^{CT} ir s_{xy}^{CIN}, kokybės metrikos (ROC ir jautrumo-specifiškumo kreivės) mokymo ir testavimo potinkliams, paruoštiems iš tinklo *SSA*. Jautrumo-specifiškumo kreivių jautrumo ašys paverstos į logaritmo pagrindu dešimt skalę.



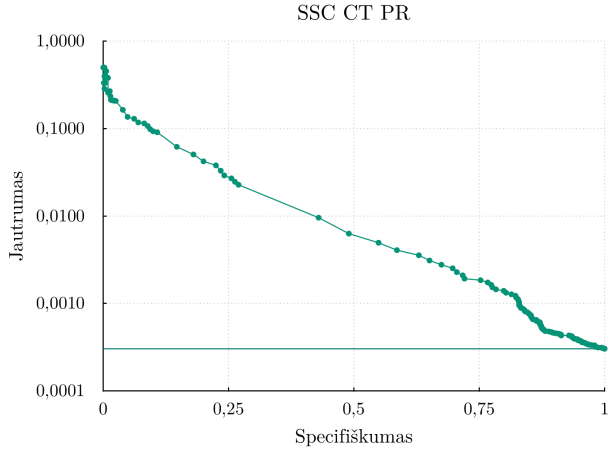
(a) s_{xy}^{CW}



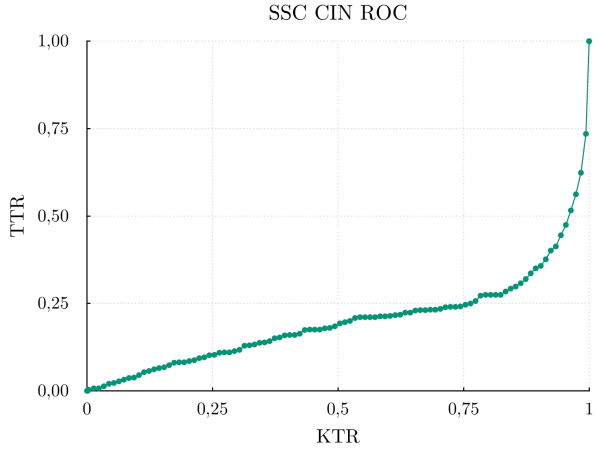
(b) s_{xy}^{CW}



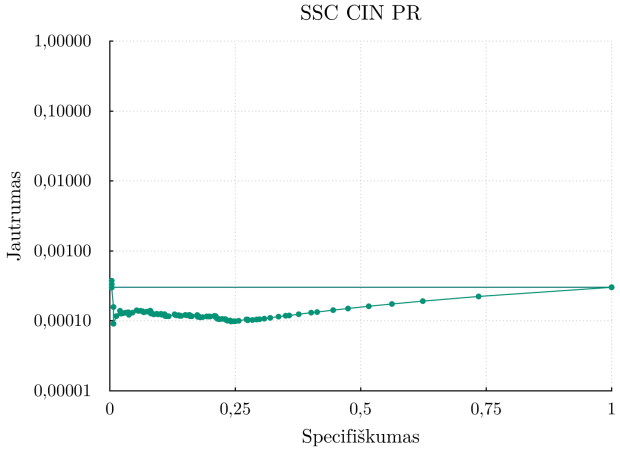
(c) s_{xy}^{CT}



(d) s_{xy}^{CT}



(e) s_{xy}^{CIN}



(f) s_{xy}^{CIN}

33 pav. Prognozės, naudojantis indeksais s_{xy}^{CW}, s_{xy}^{CT} ir s_{xy}^{CIN}, kokybės metrikos (ROC ir jautrumo-specifiškumo kreivės) mokymo ir testavimo potinkliams, paruoštiems iš tinklo *SSC*. Jautrumo-specifiškumo kreivių jautrumo ašys paverstos į logaritmo pagrindu dešimt skalę.