Graafien louhinta ja keskeiset solmut sosiaalisten verkostojen yhteisöjen analysoinnissa

Marjo Nopola

Tampereen yliopisto
Informaatiotieteiden yksikkö
Tietojenkäsittelyoppi
Pro gradu -tutkielma
Ohjaaja: Kati Iltanen
Kesäkuu 2014

Tampereen yliopisto
Informaatiotieteiden yksikkö
Tietojenkäsittelyoppi
Marjo Nopola: Graafien louhinta ja keskeiset solmut sosiaalisten verkostojen yhteisöjen analysoinnissa
Pro gradu -tutkielma, 87 sivua
Kesäkuu 2014

Sosiaalisista verkostoista voidaan löytää yhteisöjä ja tarkastella keskeisimpiä solmuja eri mittareiden perusteella. Tässä tutkimuksessa analysoitiin YouTuben videoverkostoja, jotka haettiin kuudesta erilaisesta aiheesta. Videoiden väliseksi yhteystyypiksi valittiin yhteinen kommentoija. Tutkimuksessa tarkasteltiin, löytyykö videodatasta muodostetuista graafeista yhteisöjä ja millaisia rakenteita graafeissa esiintyy. Erityisenä tutkimuskysymyksenä oli selvittää, mitkä ovat verkostojen keskeisimmät solmut eri mittareilla tarkasteltuina ja vastaavatko nämä toisiaan. Rakenteellisina keskeisyysmittareina käytettiin astetta, painotettua astetta, ominaisvektoria, läheisyyttä ja välillisyyttä. Tarkasteluun otettiin mukaan myös katselukertojen määrä videoiden attribuuttidatasta. Eri mittarien tuottamille ”parhaille” solmuille laskettiin myös keskeisyyspisteet, jotka yhteenlaskemalla saatiin jokaisesta verkostosta selville keskeisimmät solmut.

Tutkimustuloksista huomattiin, että videoverkostoissa esiintyi yhtä aihetta lukuun ottamatta yhteisöjä. Graafeissa oli myös rakenteeltaan samanlaisia elementtejä, kuten yksi suhteellisen iso komponentti muihin graafin komponentteihin verrattuna. Lisäksi jokaisessa graafissa oli lukuisia irrallisia solmuja. Rakenteelliset keskeisyysarvot ja katselukerrat eivät vaikuttaneet korreloivan keskenään. Katselluimmat videot siis saattavat saavuttaa suuren yleisön huomion, mutta kokonaiskeskeisyyspisteiden perusteella voidaan verkostoista löytää pienempien yhteisöjen keskeisiä solmuja, joilla voi olla huomattavaa vaikutusvaltaa. Tällaiset solmut voisivat vain yhdellä keskeisyysmittarilla tarkasteltuna jäädä havaitsematta.

Avainsanat ja -sanonnat: graafien louhinta, yhteisöjen louhinta, sosiaalisten verkostojen analysointi, keskeiset solmut, keskeisyysmittarit.

Sisällysluettelo

1. Johdanto

Internetin leviäminen ja verkkosisällön räjähdysmäinen kasvu ovat luoneet tarpeita uudenlaisille tiedon analysoinnin lähestymistavoille. Internetin sosiaalisten verkostojen ja sosiaalisen median sovellusten määrän lisääntyminen on merkinnyt myös uudenlaisten verkkoyhteisöjen syntymistä. Sosiaalisten verkostojen kompleksisen sisällön analysointia varten tarvitaan keinoja datassa esiintyvien suhteiden havainnollistamiseen ja erilaisia mittareita verkostojen ominaisuuksien laskemiseen. Tällaisen datan esittämiseen käytetään usein graafeja, joita tutkimalla voidaan saada merkittävää informaatiota verkostojen ja niihin sisältyvien pienempien alirakenteiden ominaisuuksista. Sosiaalisten verkostojen analysointiin voidaan käyttää eräänä menetelmänä *graafien louhintaa* (graph mining), jonka avulla verkostojen ominaisuuksia voidaan mitata ja tarkastella. Suurista verkostoista voidaan etsiä ja tarkastella pienempiä osakokonaisuuksia ja elementtejä, kuten yhteisöjä ja yksittäisiä solmuja. Graafien louhinnasta yhteisöjen analysointia varten käytetään myös nimitystä *yhteisöjen louhinta* (community mining). Louhintamenetelmien avulla voidaan selvittää verkostojen yleistä rakennetta ja löytää piileviä sääntöjä yhteisöjen evoluution ominaispiirteistä [Duan *et al.*, 2009]. Yhteisöistä voidaan löytää myös sisäkkäisiä ja hierarkkisia rakenteita.

Yhteisöjen analysointia voidaan jatkaa aina yksityiskohtaiselle solmutasolle asti. Esimerkiksi markkinoinnissa, tuotekehittelyssä ja trendien tutkimisessa on tärkeää löytää keskeisimpiä toimijoita verkostoista ja yhteisöistä. Viraalimarkkinoinnin parissa on hyödyllistä löytää esimerkiksi sellaisia toimijoita, jotka ovat tiiviisti muiden verkoston toimijoiden kanssa yhteydessä, koska tällöin voidaan saavuttaa suurin mahdollinen yleisö jonkin tuotteen, tuotemerkin tai kampanjan mainostamiselle [Landherr *et al.*, 2010]. Keskeisiä toimijoita voidaan etsiä verkostojen rakennetta tutkimalla ja määrittämällä yhteisöistä keskeisiä solmuja ja särmiä. Sosiaalisten verkostojen keskeisien solmujen määrittelyä varten on kehitetty useita erilaisia keskeisyysmittareita, jotka kuvaavat solmujen keskeisyyttä hieman eri näkökulmista.

Tässä tutkielmassa perehdytään sosiaalisten verkostojen yhteisöjen ja pienempien alirakenteiden analysointiin erityisesti sosiaalisen median sovellusten näkökulmasta. Verkostojen ominaisuuksia ja erilaisia mittareita esitellään graafiteorian käsitteiden avulla. Tutkielman tutkimusosuudessa analysoidaan eriai-

heisia YouTuben videoverkostoja ja selvitetään, löytyykö videoverkostoista samankaltaisia rakenteita ja yhteisöjä. Jokaisesta verkostosta pyritään etsimään keskeisimpiä videoita eri keskeisyysmittareiden perusteella. Videoita edustaville solmuille lasketaan kokonaiskeskeisyyspisteet keskeisyysmittareiden ja sovelluskohtaisen attribuutin eli videon katselukertojen perusteella. Vastaavanlaista useaa keskeisyysmittaria hyödyntävää keskeisyyspisteiden laskutapaa ei ole havaittu aiemmissa tutkimuksissa. Lisäksi YouTubeen liittyviä keskeisyystutkimuksia on raportoitu varsin vähän. Tässä tutkielmassa uutena esitellyn kokonaiskeskeisyyspistemäärän perusteella jokaisesta videoverkostosta määritellään keskeisimmät solmut ja esitetään ne myös visuaalisesti graafissa. Mielenkiinnon kohteena on erityisesti kysymys, vastaavatko keskeisyysmittareiden perusteella tärkeimmät solmut katselukerroiltaan suurimpia solmuja.

Luvussa 2 esitellään sosiaalisten verkostojen analysointia ja menetelmien käyttökohteita sekä hahmotetaan sosiaalisten verkostojen yhteisöjen käsitettä. Luvun lopussa esitellään sosiaalisen median palvelujen tutkimuskohteita. Luvussa 3 esitellään yhteisöjen ja verkostojen kuvaamiseen käytettävää graafia ja graafiteoriaan liittyviä keskeisiä käsitteitä sekä erilaisia graafien visualisointi- ja esittämistapoja. Luvussa 4 paneudutaan yhteisön eri määrittelytapoihin graafiteoriaan perustuen. Luvussa 5 perehdytään keskeisimpiin graafien klusterointimenetelmiin yhteisöjen löytämisen näkökulmasta. Luvussa 6 keskitytään yhteisöjen analysointiin, merkittävimpiin yhteisöjä kuvaaviin mittareihin ja dynaamisiin yhteisöihin. Luvussa 7 esitellään solmujen ja särmien analysointia sekä merkittävimpiä solmuihin ja särmiin liittyviä mittareita erityisesti solmujen keskeisyysmittareihin painottuen. Luvussa 8 esitellään keskeisyysmittareihin ja YouTuben verkostoihin liittyviä aikaisempia tutkimuksia. Tämän jälkeen luvussa 9 esitellään YouTuben videodatasta tehtyä tutkimusta, jossa analysoitiin viiden eriaiheisen videoverkoston rakennetta ja erityisesti verkostoista löytyviä keskeisiä solmuja. Luvussa 10 jatketaan pohdintaa löydettyjen videoverkostojen rakenteesta, yhteisöllisyydestä ja keskeisiin solmuihin liittyvistä tuloksista, sekä esitellään ideoita mielenkiintoisista jatkokehitysmahdollisuuksista tutkimukseen liittyen. Lopuksi luvussa 11 on yhteenveto.

2. Sosiaalisten verkostojen analysointi

Internet itsessään on esimerkki sosiaalisesta verkostosta. Tutkimus sosiaalisten verkostojen parissa on kuitenkin alkanut jo ennen internetin kehittämistä. Sosiaaliset verkostot tarkoittavat mitä tahansa ihmisten ja asioiden välisten yhteyksien kerääntymää, jotka ennen teknologian kehitystä olivat useimmiten näkymättömiä [Hansen *et al.*, 2011]. *Sosiaalisten verkostojen analysointi* (social network analysis) alkoi 1930-luvulla mm. antropologi Alfred Radcliffe-Brownin [Scott, 2000] ja ryhmäpsykoterapian pioneeri Jacob Morenon tutkimusryhmien [Hansen *et al.*, 2011] kehittämänä ja jatkui 1970-luvulle asti melko epäteknisenä, johon asti tutkittiin lähinnä sosiaalista rakennetta toimintoineen, esimerkiksi ryhmädynamiikkaa. Bavelas (1948) tutki erityisesti pienissä ryhmissä tapahtuvaa kommunikointia ja esitti oletuksen, että rakenteellisen keskeisyyden ja vaikutusvallan välillä on yhteys. Useat muutkin tutkijat jatkoivat keskeisyystutkimuksia, joissa todettiin keskeisyyden yhteys ryhmän tehokkuuteen ongelmatilanteissa, osallistujien henkilökohtaiseen tyytyväisyyteen ja käsitykseen johtajuudesta. Tutkimukset laajenivat 1960- ja 1970-luvuilla koskemaan yritysten välistä keskeisyyttä. Esimerkiksi Beauchamp (1965) esitti keskeisyyden hyödyntämistä kahden tai useamman organisaation fuusiossa. Organisaatiot voisi tällöin yhdistää niiden yksiköiden keskeisimmistä pisteistä, mikä parantaisi uuden organisaation tehokkuutta. Rogers (1974) puolestaan tutki yritysten välisiä suhteita ja huomasi, että verkoston sisäisen organisaation keskeisyys voitiin ennustaa sen ominaispiirteiden ja koko verkoston ominaisuuksien perusteella. [Freeman, 1979.] Ennen teknologian kehitystä ja erityissovellusten käyttömahdollisuutta tutkittiin lisäksi jo verkostojen kirjoa ja tiheyttä [Scott, 2000].

Miljardit ihmiset ovat 1990-luvulta lähtien liittyneet virtuaalisiin sosiaalisiin verkostoihin ja käyttäneet erilaisia sosiaalisen median sovelluksia. Sosiaalisella medialla tarkoitetaan verkkosovelluksia ja -työkaluja, jotka tukevat käyttäjien välistä sosiaalista vuorovaikutusta ja kanssakäymistä [Hansen *et al.*, 2011]. Tällaisten verkkosovellusten, -työkalujen ja -palvelujen kirjo on laaja: sosiaalinen media on erilaisia sosiaalisten verkostojen sivustoja, blogeja, mikroblogeja, pikaviestin- ja keskustelupalveluja, foorumeita, kuvien- ja videoidenvälityssovelluksia, wikejä sekä peliyhteisöjä sähköpostiohjelmia unohtamatta. Sosiaalista mediaa käytetään perheenjäsenten, ystävien ja työtovereiden yhteydenpitoon, mutta yhtä hyvin yritysten ja erilaisten instituutioiden tiedonvälitykseen ja markkinointiin. Suosittuja sosiaalisen median sovelluksia, kuten mikroblogi-

palvelu Twitteriä ja yhteisöpalvelu Facebookia, siteerataan myös uutisissa ja käytetään esimerkiksi poliittisiin vaikuttamistarkoituksiin. Kehityksen myötä myös sosiaalisten verkostojen analysointi on muuttunut, ja uudenlaisia työkaluja on kehitetty sosiaalisen datan keräämiseen, analysointiin ja visualisointiin.

Verkostojen tutkiminen on haastavaa, koska ne ovat useimmiten kooltaan suuria ja dynaamisia. Verkostoissa saattaa olla miljoonia tai jopa miljardeja solmuja [Webb and Copsey, 2011]. Yhteisöjen löytäminen on tärkeä tehtävä tällaisten suurten verkostojen analysoinnissa [Papadopoulos *et al.*, 2011]. Löydettyjen yhteisöjen rakennetta ja ominaisuuksia sekä evoluutiota voidaan analysoida. Sosiaalisen median yhteisöjä analysoimalla voidaan saavuttaa hyödyllistä tietämystä maailman eri ilmiöistä, mutta yhtä hyvin yhteisöjä saatetaan tutkia esimerkiksi verkkosisällön personoimista, suosittelujärjestelmiä ja markkinointia varten. Sosiaalisia verkostoja voidaan analysoida myös sosiologian, epidemiologian ja kriminologian tarpeisiin [Takaffoli *et al.*, 2011]. Sosiaalisten verkostojen kautta voidaan jakaa julkista tietoa liittyen esimerkiksi säävaroituksiin tai tartuntatautien leviämiseen.

Scottin [2000] mukaan sosiaalisten verkostojen analysointi soveltuu parhaiten *relaatiodatan* (relational data) tutkimiseen. Tällaisella datalla tarkoitetaan erilaisia suhdemuuttujia, jotka yhdistävät eri *toimijoiden* (agent) pareja muodostaen laajempia relaatiojärjestelmiä. Suhdemuuttuja voi merkitä toimijoiden välistä yhteyttä, sidettä, tapaamista tai muuta kontaktia. Merkittävää siis on, että relaatiota ei voi sitoa yhden toimijan ominaisuudeksi, vaan sillä kuvataan kahden tai useamman tekijän välistä yhteyttä. Relaatiodata voi esimerkiksi kuvata, kuka on kenenkin ystävä tai ketkä pitävät samoista asioista. Nämä yhteydet tarkoittavat linkkejä, joita tutkitaan verkostoanalyysin eri menetelmillä. Sosiaalisten verkostojen analysointi pyrkii selittämään linkittyneiden henkilöiden välistä sosiaalista käyttäytymistä kvantitatiivisten mittareiden avulla [Bodendorf and Kaiser, 2010].

Bodendorfin ja Kaiserin [2010] mukaan sosiaalisten verkostojen analysointi voidaan jakaa yksilötasolla tehtävään ja kollektiiviseen tutkimukseen. Yksilötasolla tutkitaan yksittäisten käyttäjien tottumuksia, mutta kollektiivinen analysointi ottaa huomioon koko verkoston toiminnan ja rakenteen.

2.1. Yhteisöt sosiaalisissa verkostoissa

Yhteisön käsitteelle ei ole yksiselitteistä määritelmää, vaan siihen vaikuttavat yleensä joko järjestelmän verkoston laajempi rakenne tai joidenkin objektien

paikalliset ominaisuudet [Papadopoulos *et al.*, 2011]. Lisäksi yhteisön käsitteen määrittelyyn vaikuttaa tutkimuksen kohteena oleva sovellusalue. Yhteisön käsite voi vaihdella myös saman sovellusalueen eri verkostojen välillä [Chen *et al.*, 2009].

Han ja Kamber [2006] määrittelevät yhteisön ryhmänä objekteja, joilla on joitakin yhteisiä ominaisuuksia. Sosiaalisten verkostojen sovelluksissa on lukuisia objekteja, jotka voivat olla monin eri tavoin yhteydessä toisiinsa erilaisten suhteiden ja vuorovaikutusmuotojen kautta. Objekteilla tarkoitetaan sovellusten erityyppisiä toimijoita, kuten käyttäjiä, sisällön kohteita (esim. valokuvat, videot ja kirjoitukset) sekä metadataan liittyviä kohteita (esim. tunnisteet ja aiheluokittelut) [Papadopoulos *et al.*, 2011]. Fortunaton ja Castellanon [2007] mukaan yhteisö on ryhmä solmuja, joilla voi yhteisten ominaisuuksien lisäksi tai niiden sijaan olla samankaltainen rooli graafissa. Verkostojen objekteja kutsutaan siis toimijoiksi, solmuiksi ja joissakin yhteyksissä myös ihmisiksi tai käyttäjiksi. Objektien välisiä yhteyksiä puolestaan kutsutaan särmiksi, siteiksi tai linkeiksi kontekstista riippuen. Solmuihin ja graafeihin liittyviin käsitteisiin perehdytään tarkemmin luvussa 3 ja graafiteorian pohjalta tehtäviin yhteisöjen erilaisiin määrittelytapoihin luvussa 4. Solmujen erilaisia rooleja esitellään kohdassa 7.1.

Erilaisia verkostoja ja yhteisöjä voidaan löytää esimerkiksi biologian, sosiaalitieteiden ja tietojenkäsittelytieteen alueilta. Kirjallisuudessa esiintyvä yhteisön käsite ei aina välttämättä ole kontekstiltaan sosiaalinen ja ihmisten välinen. Yhteisö voi siis tarkoittaa vaikkapa solun samankaltaisia proteiineja tai internetin samasta aiheesta kertovia verkkosivuja, mutta tässä tutkielmassa yhteisön käsitteellä tarkoitetaan internetin sosiaalisten verkostojen ja sosiaalisen median sisältämiä ryhmiä. Kyseisten verkostojen toimijoina ovat ihmiset, jotka ovat yhteydessä jollakin tavalla toisiinsa ainakin virtuaalisesti.

Yksilöiden välisiä verkostoja voidaan kutsua myös *pienen maailman (sosiaalisiksi) verkostoiksi* (small world (social) networks). Käsite kuvastaa sitä, kuinka kahdella toisilleen tuntemattomalla henkilöllä voi olla yhteyksiä heille yhteisten tuttavien kautta. Tällöin voidaan päivitellä sitä, kuinka ”pieni maailma onkaan”. [Han and Kamber, 2006.] *Pienen maailman efekti* (small world effect) johtaa myös tiedon nopeaan leviämiseen sekä positiivisessa että negatiivisessa mielessä – tärkeät uutiset leviävät ripeästi, mutta myös verkkovirukset ja -huijaukset. Sosiaalisten verkostojen toimijat ovat ihmisiä tai ihmisryhmiä, jotka ovat yhteydessä toistensa kanssa perhesiteiden, ystävyyssuhteiden ja työsuhteiden kautta.

[Webb and Copsey, 2011.] Yhtä hyvin internetin verkkokaupat ja monet sosiaalisen median sovellukset muodostavat erilaisia verkostoja ja yhteisöjä. Lisäksi yhteisö saattaa muodostua jonkin yhteisen mielenkiinnon tai harrastuksen ympärille.

2.2. Sosiaalisen median analysointi

Hansen ja muut [2011] jakavat sosiaalisen median palvelujen tutkimuskohteiden ominaisuudet kuuteen päätekijään:

1. *Tuottaja- ja kuluttajapopulaation koko*. Useimmissa sosiaalisen median sovelluksissa tuottajat ja kuluttajat kuuluvat samaan käyttäjäryhmään siten, että samat käyttäjät toimivat sekä sisällön tuottajina että kuluttajina eri hetkinä. Esimerkiksi YouTube-videopalvelun käyttäjät saattavat sekä tuottaa sivustolle videosisältöä että kuluttaa eli katsella muiden lataamia videoita. Populaation koko kuitenkin vaihtelee eri palveluissa, esimerkiksi sähköpostia kirjoittaa yleensä vain yksi henkilö, mutta wikidokumenttia saattaa tuottaa useampi käyttäjä. Samoin sisällön kuluttajien määrä voi vaihdella yhdestä useisiin. Monissa tunnetuissa sosiaalisen median sivustoissa sekä tuottajien että kuluttajien määrä on suuri.

2. *Vuorovaikutuksen tahti*. Perinteisesti vuorovaikutustavat on jaettu asynkroniseen ja synkroniseen. Sähköposti ja keskustelufoorumit ovat esimerkkejä asynkronisesta kommunikoinnista, jossa käyttäjät voivat osallistua vuorovaikutukseen oman aikataulunsa mukaisesti, mikä mahdollistaa sujuvamman kommunikoinnin eri aikavyöhykkeiden välillä. Lisäksi tuotettu sisältö on harkitumpaa. Synkronisessa vuorovaikutuksessa, kuten pikaviestinpalveluissa, kommunikointi käyttäjien välillä on samanaikaista. Vuorovaikutuksen tahti vaikuttaa siihen, millaisia ryhmiä eri sovellusten pariin syntyy. Jossain määrin tämä perinteinen jako on kuitenkin hämärtynyt. Esimerkiksi Facebook-käyttäjän statuspäivitykseen tai YouTube-videoon saatetaan lisätä kommentti heti sen ilmestyessä tai vasta paljonkin myöhemmin.

3. *Peruselementtien tyyppi*. Sosiaalisen median järjestelmien peruselementeillä tarkoitetaan erityyppisiä ja -kokoisia digitaalisia objekteja. Elementit voivat olla esimerkiksi videoita, valokuvia, käyttäjiä, viestejä tai erilaisia kirjoituksia. Koko voi tarkoittaa vaikkapa kirjoituksen pituutta, joka esimerkiksi *twiiteissä* (tweet) eli Twitterin viesteissä on rajoitettu 140 merkkiin. Toisaalta joissakin sosiaalisen median järjestelmissä on

lukuisia erilaisia peruselementtejä, esimerkiksi Facebook sisältää profiilisivuja, viestejä, sovelluksia, valokuvia, tunnisteita, ryhmiä jne. YouTube puolestaan sisältää eri kategorioihin kuuluvia videoita, käyttäjien luomia kanavasivuja, kommentteja, yksityisviestejä, tilaajia jne. Peruselementtien tunnistaminen on kuitenkin tärkeää, koska ne toimivat vuorovaikutuksen, ja koko verkoston, perustana.

4. *Peruselementtien kontrollointi*. Rajoituksilla voi olla vaikutuksia siihen, millaisiin vuorovaikutussuhteisiin käyttäjät haluavat osallistua. Peruselementtejä kontrolloidaan ja rajoitetaan eri tavoin sisällön luomiseen, muokkaamiseen, lukemiseen ja jakamiseen sekä moniin muihin toimintoihin liittyen. Joissakin sovelluksissa rajoitukset vaihtelevat käyttäjien roolien mukaan, esimerkiksi rekisteröityneillä käyttäjillä voi olla erilaiset oikeudet kuin anonyymeilla käyttäjillä. Esimerkiksi YouTube-videon kommentointi ja arviointi edellyttävät sivustolle rekisteröitymistä ja kirjautumista. Tällaisilla rajoituksilla voidaan pyrkiä vähentämään vahingollista sisältöä sivustoilla. Peruselementtien kontrollointi vaikuttaa siihen, miten avoin yhteisö on. Liian avoin yhteisö saattaa kärsiä vahingollisesta ”roskasisällöstä”, mutta toisaalta liika kontrollointi voi vähentää yhteisön myötävaikuttajien määrää. Tietynlaiset rajoitukset kuitenkin edistävät ryhmän jäsenten välistä säännöllistä vuorovaikutusta. Elementtien kontrollointi saattaa vaihdella järjestelmän eri käyttäjäryhmien sisällä.

5. *Yhteystyypit.* Verkostojen rakentamiseksi ja ymmärtämiseksi tulee tietää, millaisia yhteyksiä ja sidoksia elementtien välillä on yhteisöissä. Näiden yhteyksien kokoelmat muodostavat laajoja sosiaalisia järjestelmiä, joita voidaan sitten analysoida erilaisten mittareiden ja työkalujen avulla. Sosiaalisen median yhteydet voidaan jakaa implisiittisiin ja eksplisiittisiin. Ystävystyminen on ehkä yleisin eksplisiittinen yhteystyyppi sosiaalisissa verkostoissa. Lisäksi esimerkiksi YouTubessa videokanavan tilaaminen on eksplisiittinen yhteystyyppi. Käyttäjät luovat eksplisiittiset yhteydet tietoisesti, mutta implisiittiset muodostuvat verkkokäyttäytymisen perusteella. Implisiittiset yhteisöt esiintyvät valmiina järjestelmässä ja ikään kuin odottavat löytäjäänsä ilman, että niitä tarvitsee erikseen luoda [Papadopoulos *et al.*, 2011]. Käyttäjillä saattaa olla implisiittinen yhteys toisiin käyttäjiin vaikkapa kuulumalla samaan Facebook-ryhmään tai keskustelufoorumiin, jolloin käyttäjät eivät ehkä tunne toisiaan, mutta heillä todennäköisesti on samoja kiinnostuksen-

kohteita tai harrastuksia. Myös kun käyttäjät kommentoivat samaa videota YouTubessa, muodostuu implisiittinen yhteys käyttäjien välille. Käyttäjät myös jättävät järjestelmiin dataa, joka ei ole julkista, mutta jota voidaan analysoida ja josta voidaan muodostaa malleja. Tällaisia ovat esimerkiksi lukemistottumukset keskustelufoorumeilla tai käyttäjien sijaintitiedot.

Yhteydet voidaan jakaa myös suunnattuihin ja suuntaamattomiin. Käyttäjien yhdessä luoma yhteys on suuntaamaton, mutta esimerkiksi jonkun henkilön seuraaminen Twitterissä tai videokanavan tilaaminen YouTubessa luovat suunnatun yhteyden, koska tällöin päätöksen yhteydestä voi luoda vain toinen käyttäjistä. Yhteyden suunnalla on tällöin merkitystä.

Yhteyksillä on myös erilaiset merkitykset ja painotukset. Vaikkapa käyttäjien välisten viestien määrä vaihtelee, jolloin yhteyttä voidaan kuvata painoarvon avulla, joka siten kuvaa yhteyden voimakkuutta.

6. *Sisällön säilyttäminen*. Joissakin järjestelmissä, kuten wikeissä, luodaan pysyvä historia kaikista sisällön tapahtumista. YouTube säilyttää palveluun lisätyt videot pääsääntöisesti, mutta poistaa tekijänoikeuksia ja sivuston sääntöjä rikkovat videot. Toki käyttäjä itse voi poistaa lataamansa videon. Toisaalta joissakin pikaviestinsovelluksissa sisältöä eli vuorovaikutusta ei välttämättä tallenneta lainkaan. Monet sosiaalisen median sovellukset toimivat kuitenkin ääritapausten välillä, eli sisältöä saatetaan säilyttää, mutta vain tietyn aikaa. Joissakin tapauksissa säilyttäminen voi olla käyttäjän asetuksista tai tietystä tuotteesta riippuvainen.

Tutkimuskohteiden ominaisuuksien päätekijöiden selvittäminen ja ymmärtäminen on tärkeää, jotta verkostoja ja yhteisöjä voidaan analysoida oikein ja sopivien menetelmien avulla. Esimerkiksi populaation suuri koko voi rajoittaa käytettävien menetelmien valintaa laskennan mahdollisen raskauden vuoksi. Peruselementtien kontrollointi ja vuorovaikutuksen tahti puolestaan voivat vaikuttaa tuloksiin ja niiden tulkintaan, koska data voi olla rajoitusten vuoksi puutteellista tai dynaamisuuden vuoksi kovin herkkää muutoksille.

3. Verkostojen ja yhteisöjen kuvaaminen graafina

Verkostoanalyysissa ja yhteisöjen kuvaamisessa voidaan käyttää formaalina lähestymistapana graafiteoriaa. Erilaisia verkostoja voidaan mallintaa graafeina ja tutkia niiden ominaisuuksia graafianalyysin menetelmillä. Graafista voidaan käyttää myös nimitystä *sosiogrammi* (sociogram) ihmisten välisiä yhteyksiä kuvattaessa. Graafien avulla voidaan kuvata kompleksisia suhteiden joukkoja ja esittää niitä visuaalisesti erilaisten objektien avulla. Graafianalyysin avulla verkostosta voidaan laskea ja tutkia erilaisia arvoja, kuten koko, muoto ja tiheys [Hansen *et al.*, 2011]. Lisäksi graafista voidaan analysoida elementtien sijaintia verkostossa ja paikallistaa tärkeitä solmuja sekä ryhmittymiä.

3.1. Graafin peruskäsitteitä

Graafia kuvataan tyypillisesti notaatiolla $G = (V, E)$, jossa G merkitsee koko graafia eli verkostoa, V:llä kuvataan solmuja (vertices, nodes) ja E:llä solmujen välisiä *särmiä* (edges). Solmut muodostavat joukon $V = \{v_1, \ldots, v_n\}$ ja särmät joukon $E = \{e_1, \ldots, e_m\}$. Jokainen särmä voidaan merkitä solmujen parina $e_k = (v_i, v_j)$, jossa $v_i \neq v_j$ ja $v_i, v_j \in V$ [Webb and Copsey, 2011]. Solmuja voidaan sanoa myös *pisteiksi* (points) ja särmiä *viivoiksi* (lines) [Koivisto ja Niemistö, 2001]. Solmuilla kuvataan sosiaalisten verkostojen entiteettejä ja särmillä entiteettien välisiä linkkejä eli suhteita. Sosiaalisten verkostojen analysointia kutsutaan täten myös linkkianalyysiksi tai linkkien louhinnaksi [Han and Kamber, 2006]. Solmun $v_i \in V$ *vierussolmuja* (adjacent) ovat ne solmut, jotka on yhdistetty särmällä $e_k \in E$ solmuun v_i [Webb and Copsey, 2011].

Solmua voidaan kutsua myös noodiksi, toimijaksi tai entiteetiksi ja se voi edustaa monia asioita, kuten ihmistä, työryhmää, organisaatiota tai sisältöä, joka voi olla vaikkapa tunniste, valokuva tai video. Solmut voivat edustaa myös fyysisiä tai virtuaalisia paikkoja sekä tapahtumia. Attribuuttien lisääminen solmuille ei ole pakollista, mutta niiden avulla voidaan ehkä parantaa verkoston analysointia ja visualisointia. Attribuuteilla voidaan kuvata esimerkiksi ihmisen ikää, sukupuolta, rotua, sijaintia tai tietoja, jotka liittyvät enemmänkin henkilön järjestelmäkäyttäytymiseen, kuten viestien tai sisäänkirjautumisten määrää. [Hansen *et al.*, 2011.] Verkkosisältöä edustavien solmujen attribuuteilla voidaan kuvata erilaisia sisällön ominaisuuksia, esimerkiksi videon kuvailutietoja. Solmut erotellaan graafissa *leimojen* (label) avulla, jotka ovat usein erilaisia merkkijonoja.

Särmää voidaan kutsua linkiksi, yhteydeksi, siteeksi tai suhteeksi. Särmät voivat kuvata monenlaisia solmujen välisiä suhteita, kuten kommunikointia, sukulaisuutta, ystävyyttä, läheisyyttä, työsuhdetta, kauppakumppanuutta, sitaatteja, toimintaa tai yhteisiä attribuutteja. Särmät ovat joko *suuntaamattomia* (undirected) tai *suunnattuja* (directed), joiden mukaan myös graafeja kutsutaan suuntaamattomiksi tai suunnatuiksi. Suuntaamattoman särmän avulla kuvataan suhdetta, jossa ei ole selkeää alku- ja päätepistettä, vaan suhde on yhteinen [Hansen *et al.*, 2011]. Suunnattuun särmään on nimensä mukaisesti merkitty särmän suunta. Särmiä kutsutaan tällöin usein *kaariksi* (arcs) tai *nuoliksi* (arrows). Nuolella voidaan vaikkapa kuvata tilannetta, jossa Twitter-käyttäjä seuraa toista käyttäjää. Suunnatun graafin kahden eri solmun välillä voi olla korkeintaan kaksi kaarta eli yksi kumpaankin suuntaan. Tällainen tilanne voi tulla kyseeseen, kun esimerkiksi kaksi Twitter-käyttäjää seuraavat toisiaan. Jos solmusta v_i kulkee kaari solmuun v_j, sanotaan solmua v_i lähtösolmuksi ja solmua v_j maalisolmuksi. Suunnattua graafia kutsutaan myös *digraafiksi*.

Solmusta voi myös kulkea särmä solmuun itseensä, jolloin särmää kutsutaan *luupiksi* (loop) tai *itseissilmukaksi* (self-loop). Tällaisia särmiä kutsutaan myös *refleksiivisiksi* (reflexive) [Virtanen, 2003]. Kuvauksen kohteena voi tällöin olla esimerkiksi tilanne, jossa henkilö lähettää itselleen sähköpostiviestin muistutukseksi [Hansen *et al.*, 2011].

Suuntaamattoman graafin solmun v_i *aste* (degree) kuvaa solmuun liittyvien särmien lukumäärän, joka voidaan laskea kaavalla

$$degree(v_i) = \sum_{v_j \in V} a_{ij}, \quad v_i \neq v_j,$$

jossa v_j kuvaa kaikkia graafin muita solmuja ja a_{ij} solmujen v_i ja v_j välisten särmien lukumäärää [Opsahl *et al.*, 2010]. Luupit lisäävät solmun astetta kahdella. Suunnatuissa graafeissa voidaan erotella solmun *lähtö-* (outdegree) ja *tuloaste* (indegree). Lähtöaste kuvaa solmusta lähtevien ja tuloaste solmuun kohdistuvien kaarien lukumäärää [Scott, 2000].

3.2. Erityyppisiä graafeja ja verkostoja

Fortunaton [2010] mukaan yhteisöjen rakennetta kuvaavat graafit eivät ole säännöllisiä, vaan niissä vallitsee sekä järjestys että epäjärjestys. Tällä hän tarkoittaa särmien epäsäännöllistä jakautumista solmujen kesken. Graafit voivat kuitenkin paljastaa yhteisön rakenteesta myös järjestystä ja organisoituneisuut-

ta esimerkiksi hierarkkisuuden tai solmujen luokittelun muodossa. Sosiaalisen median verkostoja kuvaavat graafit voivat olla erikokoisia, suuntaamattomia, suunnattuja, *yksinkertaisia* (simple), *painotettuja* (weighted) tai *monisuuntaisia* (multiway) vaihdellen eri verkostojen luomisprosessin mukaisesti [Papadopoulos *et al.*, 2011]. Graafin koko määritellään yleensä solmujen ja särmien lukumääränä. Graafiteorian määritelmien mukaan yksinkertaisen graafin kahden eri solmun välillä voi olla korkeintaan yksi särmä ja solmusta ei voi olla särmää solmuun itseensä [Koivisto ja Niemistö, 2001]. Tällainen yksinkertainen graafi voi siis kuvata yhdenlaista suhdetta yhteisön objektien välillä. Esimerkiksi Facebook-käyttäjien ystävyyssuhteita voidaan kuvata yksinkertaisella graafilla. Yksinkertainen graafi on *täydellinen* (complete), jos jokaisen kahden eri solmun välillä on särmä [Koivisto ja Niemistö, 2001]. Käytännössä täydelliset graafit ovat hyvin harvinaisia jopa pienissä sosiaalisissa verkostoissa [Scott, 2000].

Painotetun graafin jokaiseen särmään on liitetty jokin luku, jota kutsutaan särmän *painoksi*. Sosiaalisissa verkostoissa painolla kuvataan yleensä yhteyden kestoa, emotionaalista intensiteettiä, läheisyyttä tai palvelujen vaihtoa [Opsahl *et al.*, 2010].

Monisuuntaisuus merkitsee graafiteorian mukaisesti *suunnattua multigraafia* (directed multigraph), jossa voi olla useita särmiä kahden eri solmun välillä, eli sen avulla voidaan mallintaa useaa erilaista suhdetyyppiä. Solmujen välisiin särmiin voidaan liittää ominaisuuksia, joilla kuvataan suhteen tyyppi. Verkostoista, joissa esiintyy monia suhdetyyppejä, käytetään myös nimitystä *moninkertaiset verkostot* (multiplex networks) [Hansen *et al.*, 2011].

Graafin solmut voivat kuulua eri luokkiin ja särmiä saattaa esiintyä vain eri luokkien edustajien välillä. Graafia, jossa solmut kuuluvat kahteen eri luokkaan, eikä saman luokan edustajien välillä ole särmää, kutsutaan *kaksijakoiseksi* (bipartite graph). [Webb and Copsey, 2011.] Kaksijakoisella graafilla voidaan vaikkapa mallintaa käyttäjien lisäämiä valokuvia, jolloin käyttäjät ja valokuvat kuuluvat eri luokkiin siten, ettei yhdenkään käyttäjä- tai valokuvaparin välillä kulje särmää. Tällaisia yksilöiden ja jonkun tapahtuman, toiminnon tai sisällön välisiä suhteita kutsutaan myös *yhteysverkostoksi* (affiliation network), jota käytetään monien suosittelujärjestelmien materiaalina [Hansen *et al.*, 2011]. Täsmälleen kahden solmutyypin verkostoa sanotaan *bimodaaliseksi* (bimodal network) [Hansen *et al.*, 2011].

Monijakoiset (multipartite) graafit mallintavat useampaa kuin kahta solmuluokkaa ja niiden välisiä yhteyksiä. Puhutaan myös *multimodaalisista verkostoista* (multimodal networks), kun verkostoissa on erityyppisiä solmuja [Hansen *et al.*, 2011]. Kaksi- ja monijakoisista graafeista voidaan mallintaa erilaisia kuvauksia, joihin otetaan mukaan vain tietty luokka tai luokkapari [Fortunato, 2010]. Vain yhdentyyppisiä entiteettejä sisältäviä verkostoja kutsutaan *unimodaalisiksi* (unimodal networks). Tällaisilla verkostoilla voidaan kuvata esimerkiksi pelkästään käyttäjien välisiä yhteyksiä. Bimodaaliset yhteysverkostot voidaan muuntaa kahdeksi erilliseksi unimodaaliseksi verkostoksi. [Hansen *et al.*, 2011.]

Yhteisöjen analysoimista varten graafeissa ei useinkaan voida kuvata koko kompleksista verkostoa kaikkine vuorovaikutuksineen, vaan yhteisöstä valitaan tietyt mielenkiinnon kohteena olevat näkökulmat. Tällöin koko verkostosta otetaan analysoitavaksi osittainen verkosto. Mielenkiinnon kohteena saattaa olla tietty käyttäjäjoukko, aihealue, ajanjakso, ominaisuudet jne. Erilaisia suhteita voidaan mallintaa erityyppisten graafien avulla. Yksinkertaistetuissa graafeissa kuvataan usein vain yksi tai kaksi erilaista solmutyyppiä ja särmät ovat yksinkertaisia. Objektien välillä mallinnetaan tällöin vain yhdenlainen suhdetyyppi, jolloin graafi on *homogeeninen*. Han ja Kamber [2006] kuitenkin huomauttavat, että vain yhdenlaisen suhteen tutkiminen yhteisöistä saattaa estää hyödyllisen piilossa olevan tiedon löytämistä. Toisaalta myös suhdetyyppi saattaa olla piilossa, joten se tulee ensin paikallistaa, jotta voidaan löytää kyseisen suhteen sisältävä yhteisö [Cai *et al.*, 2005]. Todellisuudessa sosiaaliset verkostot ovat aina *heterogeenisia* ja niissä vallitsee erilaisia suhteita. Tällöin jokainen suhdetyyppi voidaan mallintaa omana suhdeverkostonaan ja sosiaalista verkostoa kutsua myös *monisuhteiseksi* (multi-relational) [Cai *et al.*, 2005].

3.3. Graafissa kulkeminen

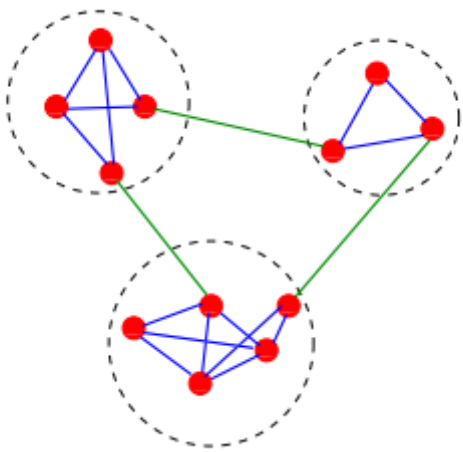
Fortunaton [2010] mukaan *yhtenäisyys* (connectedness) on yhteisön pakollinen ominaisuus. Yhteisöä kuvaavan graafin tulee siten olla yhtenäinen niin, että jokaisesta solmusta pääsee *polkua* (path) pitkin kaikkiin graafin muihin solmuihin. Polku voidaan merkitä äärellisenä jonona, jossa on vuorotellen solmuja ja särmiä [Koivisto ja Niemistö, 2001]. Polun pituus voidaan määritellä laskemalla yhteen sillä esiintyvien särmien lukumäärä. Kahden solmun välinen *geodeesinen etäisyys* (geodesic distance) on niitä yhdistävä lyhin polku. [Scott, 2000.] Geodeesisesta etäisyydestä voidaan myös käyttää pelkästään ilmausta *etäisyys* (distance) tai *geodeesi* (geodesic). Suunnatuissa graafeissa tulee huomioida, että polulla voidaan kulkea vain särmien suuntaisesti. Täten suunnatun graafin sol-

mujen v_i ja v_j välinen etäisyys ei välttämättä ole sama kuin etäisyys toiseen suuntaan solmusta v_j solmuun v_i [Opsahl *et al.*, 2010]. Yhtenäisen *graafin halkaisija* (graph diameter) on suurin geodeesinen etäisyys kahden solmun välillä [Webb and Copsey, 2011].

Polun lisäksi graafissa liikkumista voidaan kuvata erilaisten *reittien* (trail) ja *kulkujen* (walk) avulla. Reitissä jokaista särmää käytetään korkeintaan kerran, mutta samassa solmussa voidaan vierailla useasti. Kuluissa solmut ja särmät voivat esiintyä satunnaisesti ja rajoittamattomasti. Kaikki reitit ovat kulkuja, mutta kaikki kulut eivät ole reittejä, koska samaa särmää pitkin saatetaan kulkea monta kertaa. [Borgatti, 2005.]

3.4. Graafin ja verkoston alirakenteita

Verkostoa kuvaava graafi voidaan jakaa osiin eli klustereihin. Tällöin yksi klusteri kuvaa yhtä melko itsenäistä yhteisöä verkostossa. Klusterin, yhteisön ja *aligraafin* (subgraph) voidaan siten käsittää merkitsevän samaa kohdetta. Aligraafi on graafin osa, joka itsessäänkin on graafi [Koivisto ja Niemistö, 2001]. Aligraafi on siis mikä tahansa koko graafista valittu solmujen joukko, jotka yhdistyvät toisiinsa särmien avulla [Scott, 2000]. Tästä voidaan myös käyttää nimitystä *moduuli* (module), ryhmä tai luokka kontekstista riippuen [Radicchi *et al.*, 2004]. Saman klusterin sisällä on paljon solmuja yhdistäviä särmiä, mutta klusterista lähtee verrattain vähän särmiä muiden klustereiden solmuihin. Klusterin sisältämillä solmuilla on todennäköisesti yhteisiä piirteitä tai samankaltaisia rooleja. [Fortunato, 2010.] Kuvassa 1 on havainnollistettu kolme klusteria sisältävä yksinkertainen graafi, jossa yhteisöt on ympyröity katkoviivoilla.



Kuva 1. Kolmen yhteisön yksinkertainen graafi [Fortunato and Castellano, 2007].

Sosiaalisten verkostojen analyysissa käytettävä graafisanasto poikkeaa hieman graafiteoriasta. Esimerkiksi *klikki* (clique) tarkoittaa sosiaalisissa verkostoissa täydellistä, maksimaalista aligraafia, mutta graafiteoriassa klikkeinä pidetään myös täydellisiä, ei-maksimaalisia aligraafeja [Fortunato, 2010]. Aligraafin maksimaalisuus jonkin valitun ominaisuuden suhteen tarkoittaa sitä, että siihen ei voida lisätä elementtejä menettämättä kyseistä ominaisuutta [Virtanen, 2003].

Graafin *komponentiksi* (component) kutsutaan maksimaalista yhdistyneiden solmujen joukkoa [Virtanen, 2003], joka on yksinkertaisin erilaisista aligraafeista. Komponentin kaikki solmut ovat yhteydessä toisiinsa polkujen avulla, eikä sen ulkopuolelle kulje polkuja [Scott, 2000.] *Yhtenäisessä* (connected) graafissa on vain yksi komponentti. Suunnatun graafin *heikosti yhtenäisessä* (weakly connected) komponentissa jokaisesta solmusta on polku johonkin muuhun solmuun. Suunnatun graafin komponentti on *vahvasti yhtenäinen* (strongly connected), jos sen kaikkien solmuparien välillä kulkee polku molempiin suuntiin [Mislove *et al.*, 2007]. Jos graafissa on enemmän kuin yksi komponentti, se on *epäyhtenäinen* (disconnected). Irrallinen solmu muodostaa yksinään yhden graafin komponenteista. Jokainen graafin solmu ja särmä kuuluvat täsmälleen yhteen graafin komponenttiin. [Koivisto ja Niemistö, 2001.]

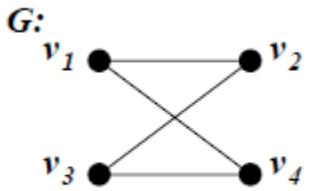
3.5. Graafien erilaisia esittämistapoja

Graafien esittämiseen voidaan käyttää erilaisia matriiseja, kuten *vierus-* (adjacency matrix) ja *tapausmatriisia* (incidence matrix). Yksinkertaisen n-solmuisen graafin G vierusmatriisi A on symmetrinen $n \times n$ -matriisi, jossa

$$a_{ij} = \begin{cases} 1, & jos\ v_i\ ja\ v_j\ ovat\ vierussolmuja, \\ 0, & muulloin. \end{cases}$$

Kuvassa 2 on yksinkertainen 4-solmuinen graafi G ja sitä vastaava vierusmatriisi A. Sekä vierusmatriisin rivit että sarakkeet kuvaavat graafin solmuja. Yksinkertaisen graafin vierusmatriisin alkiot ovat aina nollia ja ykkösiä. Multigraafia kuvaavaan vierusmatriisiin merkitään myös solmujen väliset särmien lukumäärät, jolloin alkioiden arvot luonnollisesti vaihtelevat. Myös suuntaamattoman multigraafin vierusmatriisi on symmetrinen. Lisäksi senkin lävistäjäalkiot ovat nollia, koska suuntaamattomissa graafeissa solmusta

ei voi olla särmää siihen itseensä. Solmun aste voidaan laskea vierusmatriisin solmua koskevan rivin summana.



$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

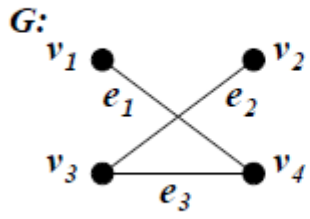
Kuva 2. Yksinkertainen graafi *G* ja sen vierusmatriisi *A* [Koivisto ja Niemistö, 2001].

Painotettu graafi voidaan esittää vierusmatriisina, johon merkitään solmujen välisen särmän painotettu arvo tai nolla, jos solmujen välillä ei ole yhteyttä [Webb and Copsey, 2011].

Suuntaamattoman n-solmuisen ja m-särmäisen graafin G tapausmatriisi T on sellainen $n \times m$ -matriisi, jossa

$$t_{ik} = \begin{cases} 1, & jos\ särmä\ e_k\ kulkee\ solmun\ v_i\ kautta, \\ 0, & muulloin. \end{cases}$$

Kuvassa 3 on esimerkki 4-solmuisesta ja 3-särmäisestä suuntaamattomasta graafista *G* sekä sitä vastaavasta tapausmatriisista *T*. Tapausmatriisin rivit vastaavat graafin solmuja *(v1-v4)* ja sarakkeet särmiä *(e1-e3)*. Tapausmatriisista voidaan siten tarkastella, mitkä solmut yhdistyvät tietyn särmän avulla, tai mitä särmiä tiettyyn solmuun yhdistyy.



$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Kuva 3. Suuntaamaton graafi *G* ja sen tapausmatriisi *T* [Koivisto ja Niemistö, 2001].

Matriisit ovat sopivia esitystapoja matemaattisia tehtäviä varten, mutta suurissa verkostoissa niitä saattaa olla hankala tutkia suuren kokonsa vuoksi. Tällöin vaihtoehtona verkoston kuvauksessa voidaan käyttää ”*solmulistaa*” (”edge list”). Binääriset verkostot voidaan kuvata merkitsemällä kahteen sarakkeeseen niiden solmujen nimet, joiden välillä kulkee särmä. Lisäsarakkeita voidaan käyttää kuvaamaan esimerkiksi painotettuja särmiä. [Hansen *et al.*, 2011.]

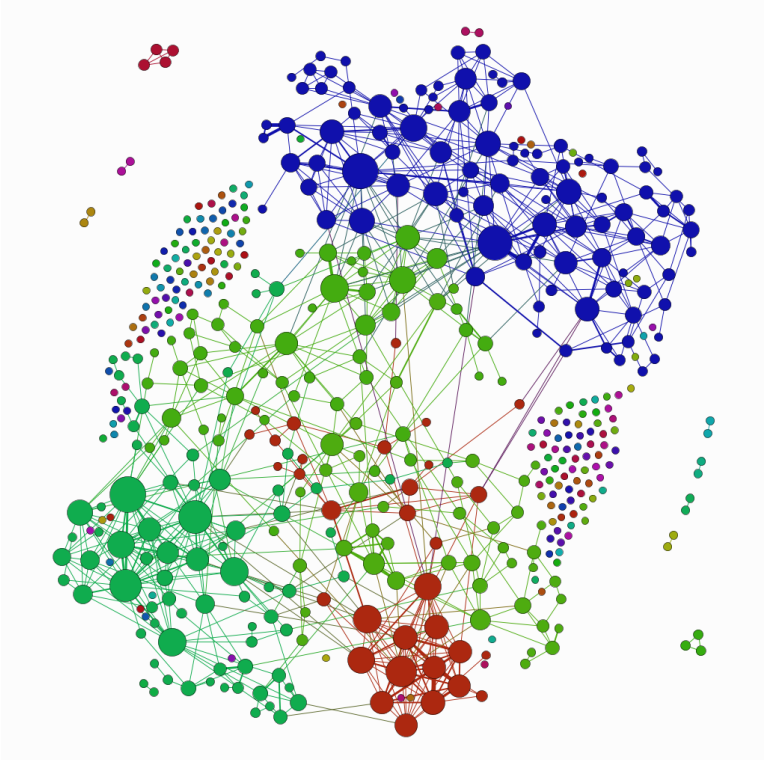
Graafien visuaaliset esitystavat havainnollistavat usein verkostossa esiintyviä suhteita ja niiden puuttumista taulukkomuotoista dataa paremmin. Graafin visualisoinnin avulla voi luoda yleissilmäyksen verkoston rakenteesta ja selventää ymmärrystä verkoston klustereista, yhteisöistä, klikeistä ja tärkeistä jäsenistä. Verkoston tärkeitä kohtia voidaan demonstroida erilaisin visuaalisin menetelmin. Haasteina graafien visualisoinneissa on kuitenkin niiden luettavuus, koska suurissa verkostoissa niistä voi helposti muodostua liian suuria ja tiheitä.

Shneiderman ehdottaa seuraavaa neljää tavoitetta verkostojen visualisoinnissa [Hansen *et al.*, 2011]:

1. Jokainen solmu on näkyvä.
2. Jokaisen solmun aste on laskettavissa.
3. Jokaisen särmän lähtö- ja maalisolmut ovat selvästi seurattavissa.
4. Klusterit ja syrjässä olevat, poikkeavat solmut ovat tunnistettavissa.

Käytännössä kaikkien tavoitteiden toteutuminen on hyvin haastavaa suurten verkostojen tapauksissa, koska solmujen ja särmien lukumäärät voivat olla hyvin suuria. Joka tapauksessa myös suurten verkostojen visualisoinnilla voidaan hahmottaa graafista yleiskuva ja havainnollistaa verkoston rakennetta, kuten komponenttien kokoa ja määrää.

Solmujen attribuuttidataa voidaan kuvata visuaalisin menetelmin, kuten solmujen koon, värin tai läpinäkyvyyden avulla. Myös erityyppiset solmut voidaan erotella erilaisilla väreillä tai muodoilla. Erikokoisilla solmuilla voidaan vaikkapa kuvata niiden tärkeyttä ja vaikutusvaltaa verkostossa jonkin valitun keskeisyysmittarin perusteella. Klusterit voidaan erotella verkostosta erivärisin solmuin, mistä nähdään esimerkki kuvassa 4. Painotettuja särmiä ilmaisevat usein vaihtelevat viivan paksuudet tai tummuudet. Eri särmätyyppejä voidaan myös kuvata erilaisilla viivoilla (esimerkiksi piste- tai kiinteät viivat) tai särmien leimoilla.



Kuva 4. Esimerkki värien käytöstä klustereiden visualisoinnissa.

4. Yhteisöjen määrittelytapoja

Sosiaalisen median yhteisö voidaan määritellä abstrakteimmillaan verkoston aligraafina, jonka toimijoilla on jokin yhteinen tekijä tai mielenkiinnon kohde. Yhteinen tekijä voi olla vaikkapa tietty aihe, henkilö, tapahtuma, paikka tai toiminto. [Papadopoulos *et al.*, 2011.] Esimerkiksi tiettyä YouTube-videota kommentoivat käyttäjät muodostavat implisiittisen yhteisön. *Yhteisöjen löytämisen* (community detection) tarkoituksena on Fortunaton [2010] mukaan yhteisöjen ja mahdollisesti niiden sisältämän hierarkkisen järjestyksen tunnistaminen graafien topologisten ominaisuuksien perusteella.

Ensimmäinen askel graafien louhinnassa on Fortunaton [2010] mukaan yhteisön kvantitatiivinen määrittely sovellusalueen mukaisesti. Hänen mukaansa useimmissa yhteisöjen määrittelyissä käytetään ohjesääntönä periaatetta, jonka mukaan yhteisön sisällä olevien särmien lukumäärän tulee olla suurempi kuin yhteisöstä muualle graafiin kulkevien särmien määrä. Webb ja Copsey [2011] jakavat implisiittisten yhteisöjen löytämisen verkostosta globaaleihin ja lokaaleihin lähestymistapoihin. Globaaleissa menetelmissä lähtökohtana on koko verkoston solmujen joukko, mutta lokaaleissa tutkimus voi alkaa yhdestä siemensolmusta ja yhteisön määrittely koskee solmujen osajoukkoa. Yhteisöjä voidaan määritellä myös muilla tavoin, kuten solmujen samankaltaisuuteen perustuen tai yhteisöjen muodostumisprosesseja tarkastelemalla. Erilaisia yhteisöjen määrittelytapoja esitellään seuraavaksi.

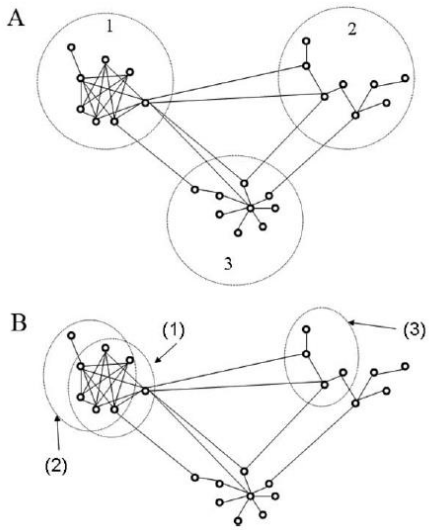
4.1. Lokaalit määrittelyt

Lokaalia määrittelyä vastaavat yhteisöt ovat useimmiten maksimaalisia aligraafeja. Tällaisiin aligraafeihin ei voida lisätä uusia solmuja tai särmiä ilman, että menetetään niitä kuvaavat ominaisuudet. Tutkittavat yhteisöt ovat siten graafin osia, joita voidaan käsitellä myös itsenäisinä entiteetteinä. Joskus tutkimusmenetelmässä otetaan myös huomioon aligraafin välitön naapurusto, mutta ei verkoston muita osia. Yhteisöllä oletetaan olevan pieni *irrotusjoukko* (cutset), eli vain vähän särmiä yhdistää sitä muuhun graafiin. [Fortunato, 2010.] Irrotusjoukolla tarkoitetaan minimaalista särmäjoukkoa, jonka poistaminen jakaa graafin kahteen tai useampaan komponenttiin [Koivisto ja Niemistö, 2001].

Klikit ovat maksimaalisia täydellisiä aligraafeja, joissa kaikki yhteisön solmut ovat yhteydessä toisiinsa. Kolmikulmaiset klikit (triangles) ovat yksinkertai-

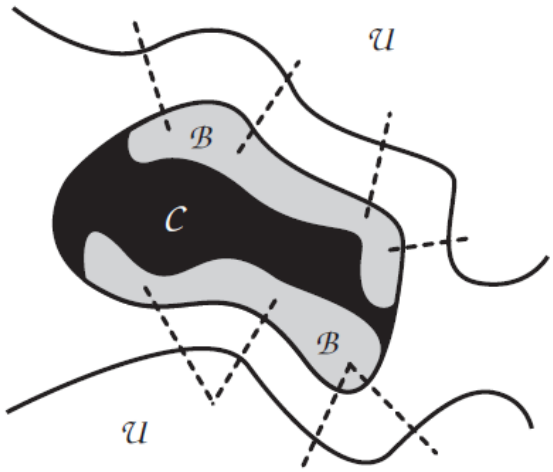
simpia ja niitä esiintyykin verkostoissa usein. Sen sijaan suuremmat klikit ovat harvinaisia. [Fortunato and Castellano, 2007.] Klikkeihin pohjautuva yhteisöjen lokaali määrittely on sellaisenaan jyrkkä, mutta sen pohjalta on kehitetty lievennettyjä yhteisöjen määrittelytapoja.

Lokaaleihin määrittelyihin kuuluu myös yhteisöjen jakaminen vahvoihin ja heikkoihin. Vahvoissa yhteisöissä jokaisella solmulla on enemmän vierussolmuja aligraafin sisällä kuin sen ulkopuolella. Täten aligraafin jokaisen solmun lähtöaste on suurempi kuin sen tuloaste. Radicchi ja muut [2004] esittelivät solmun lähtö- ja tuloasteiden käsitteet myös suuntaamattomille graafeille. Solmun lähtöaste on tällöin sellaisten särmien lukumäärä, jotka yhdistävät kyseistä solmua muihin solmuihin saman yhteisön sisällä. Solmun tuloaste sen sijaan kuvaa särmien lukumäärää, jotka yhdistävät solmua yhteisön ulkopuolisiin solmuihin. [Luo *et al.*, 2008.] Vahvan yhteisön määritelmä on Fortunaton [2010] mukaan melko jyrkkä. Myös Luo ja muut [2008] kritisoivat vahvan yhteisön määrittelyä, koska siihen saattaa voimakkaasti vaikuttaa yksittäisen solmun aste. Määrittelyä voidaan lieventää käyttämällä heikon yhteisön määritettä. Tällöin riittää, että aligraafin solmujen lähtöasteiden summa on suurempi kuin tuloasteiden summa [Radicchi *et al.*, 2004]. Kuvassa 5 on havainnollistettu yhteisön määrittelyä vahvuuden ja heikkouden määritelmien mukaisesti. Kuvan A-kohdassa verkosto on jaettu kolmeen yhteisöön *(1-3)* intuitiivisesti, mutta kohdassa B on käytetty erilaisia lähestymistapoja yhteisöjen määrittelyyn: (1) hyvin yhtenäinen yhteisö, (2) vahva yhteisö ja (3) heikko yhteisö.



Kuva 5. A) Verkosto jaettu kolmeen yhteisöön intuitiivisesti.
B) Verkosto jaettu kolmeen yhteisöön (1) yhtenäisyyden, (2) vahvuuden ja (3) heikkouden perusteella [Luo *et al.*, 2008].

Yhteisö voidaan määritellä myös silloin, kun verkostosta ei ole globaalia tietämystä. Kuvassa 6 esitetty alue *C* (Core) esittää graafin *G* sellaisten solmujen joukkoa, joiden yhteydet tunnetaan. Alueesta *U* (Unvisited) tunnetaan vain ne solmut, joiden vierussolmut ovat alueeseen *C* kuuluvalla raja-alueella *B* (Boundary). Raja-alueella *B* on siis sellaisten solmujen joukko, joilla on vähintään yksi vierussolmu alueella *U*. Tällöin graafista *G* saadaan lisätietoa vain tutkimalla näiden vierussolmujen naapurustoa alueella *U*. Tuloksena vierussolmun vierussolmusta $v_i \in U$ saadaan jäsen tunnettujen solmujen joukkoon *C*. Verkostoa pitää siten tutkia solmu kerrallaan, ja yhteisön rakennetta voidaan mitata vain globaaleista ominaisuuksista riippumattomien tekijöiden avulla.



Kuva 6. Lokaalin modulaarisuuden määrittelyssä käytetyt graafialueet [Clauset, 2005].

Tällaista lokaalia yhteisörakennetta kuvaa mittari *lokaali modulaarisuus* (local modularity) *R*, joka voidaan laskea kaavalla

$$R = \frac{\sum_{ij} B_{ij}\delta(v_i, v_j)}{\sum_{ij} B_{ij}} = \frac{I}{S},$$

jossa B_{ij} on 1, kun solmujen v_i ja v_j välillä on särmä ja jompikumpi solmuista on *B*:ssä. Muulloin B_{ij} on 0. Lisäksi $\delta(v_i, v_j)$ on 1, kun $v_i \in B$ ja $v_j \in C$ tai toisinpäin. Muulloin $\delta(v_i, v_j)$ on 0. Tässä *I* on niiden särmien lukumäärä, joiden kumpikaan

solmu ei ole *U*:ssa. *S* on niiden särmien lukumäärä, joiden yksi tai molemmat solmut ovat *B*:ssä. Clauset [2005] havaitsi tutkimuksissaan, että solmua ympäröivän yhteisön lokaali modulaarisuus korreloi negatiivisesti solmun asteen kanssa. Algoritmit, jotka kasvattavat yhteisöä solmu kerrallaan, muodostavat yhteisön hierarkian kuitenkin vain tietyn siemensolmun näkökulmasta. [Clauset, 2005.]

Luo ja muut [2008] laajensivat modulaarisuuden ja solmun asteen käsitteitä koskemaan aligraafia. Yhteisö voidaan tällöin määritellä myös *aligraafin modulaarisuuden* (subgraph modularity) perusteella. Aligraafin lähtöaste on sen sisäisten särmien lukumäärä. Tuloaste kuvaa särmien lukumäärää, joiden avulla aligraafi yhdistyy muuhun graafiin. Aligraafin modulaarisuus on siten sen lähtö- ja tuloasteen suhdeluku. Modulaarisuuden määrä kasvaa, kun aligraafilla on enemmän sisäisiä särmiä kuin ulkoisia särmiä. [Luo *et al.*, 2008.]

Yksittäiseen käyttäjään voidaan liittää lokaali *egosentrinen verkosto* (egocentric network), johon kuuluvat vain käyttäjään yhteydessä olevat muut yksilöt. Tietyn käyttäjän Facebook-ystävät muodostavat esimerkiksi egosentrisen verkoston, joka on 1-asteinen. Sen sijaan 1,5-asteiseen verkostoon kuuluvat edellisen lisäksi myös käyttäjän ystävien väliset yhteydet. Verkostoa voidaan jatkaa 2-asteiseksi, jolloin yhteisöön liitetään myös käyttäjän ystävien vastaavat yhteydet, eli ystävien ystävät. Nämä kolme verkostoa kuvaavat tietyn yksilön lokaaleja naapurustoja. [Hansen *et al.*, 2011.]

4.2. Globaalit määrittelyt

Globaalissa määrittelyssä yhteisöjä ei voida erottaa koko verkostoa kuvaavasta graafista, vaan yhteisöt ovat olennaisia osia järjestelmän toiminnassa. Yhteisörakenne on siten koko verkoston ominaisuus [Papadopoulos *et al.*, 2011]. Graafi eroaa *satunnaisesta graafista* (random graph), kun siitä voidaan erottaa yhteisön rakenne. Erdös-Rényin kehittämässä satunnaisessa graafissa jokaisella solmuparilla on sama todennäköisyys olla yhteydessä, joten siitä ei voida erottaa tiettyjä solmujen ryhmiä, eikä siinä siten ole yhteisörakennetta [Fortunato, 2010].

Papadopouloksen ja muiden [2011] mielestä ehkä paras tapa arvioida verkoston jakoa yhteisöiksi on tarkastella yhteisöjen välisten särmien määrää eli *irrotusjoukon kokoa* (cut size). Yhteisöjen välisten särmien absoluuttisen määrän käyttäminen on kuitenkin ongelmallista, joten yhteisöjaon mittarina voidaan käyttää normalisoituja arvoja, kuten *normalisoitua irrotusta* (normalized cut) ja *konduktanssia* (conductance). Normalisoitu irrotus saadaan laskemalla yhteen

eri yhteisöjen välisen erilaisuuden ja samankaltaisuuden suhdeluvut. Yhteisöjen välinen erilaisuus määritellään tässä yhteydessä laskemalla yhteisöjen irrotusjoukon koko ja samankaltaisuus laskemalla jokaisesta yhteisöstä muualle graafiin lähtevien yhteyksien määrät [Shi and Malik, 2000]. Yhteisön konduktanssi tarkoittaa todennäköisyyttä, että yhteisöstä alkavan *satunnaiskulun* (random walk) aikana poistutaan kyseisestä yhteisöstä [Gleich and Seshadhri, 2012]. Satunnaiskulku on polku, jossa suunta valitaan satunnaisesti. Alhainen konduktanssi tarkoittaa yleensä yhteisöjaon onnistumista. Vahvoissa yhteisöissä satunnaiskulkuun kuluu paljon aikaa, koska yhteisön sisäiset särmät ovat tiheässä ja siten mahdollisesti seurattavien polkujen määrä on suuri [Fortunato, 2010].

Fortunato [2010] korostaa modulaarisuuden merkitystä globaalina määrittelykriteerinä sen lisäksi, että sitä voidaan käyttää yhteisön laatumääreenä ja avaintekijänä graafien klusteroinnissa. Modulaarisuudesta on esitelty monia muunnoksia, mutta perusmääritelmän taustalla on käsite *nollamalli* (null model). Määritelmän mukaan aligraafi on yhteisö, jos sen sisäisten särmien lukumäärä ylittää särmien määrän odotusarvon, joka samalla aligraafilla olisi nollamallissa. Särmien määrän odotusarvo on kaikkien mahdollisten nollamallin toteutumien keskiarvo. Nollamallilla tarkoitetaan graafista tehtyä mallia, joka vastaa alkuperäistä graafia joidenkin rakenteellisten piirteiden osalta, mutta on muilta osin satunnainen graafi. Nollamallia käytetään apuna vertailussa, kun halutaan selvittää, onko tutkittavassa graafissa havaittavissa yhteisörakenne. [Fortunato, 2010.] Modulaarisuusarvo on aina pienempi kuin 1, ja arvo voi olla myös negatiivinen. Esimerkiksi pelkästään irrallisia solmuja sisältävän graafin modulaarisuus on negatiivinen. Vain positiiviset modulaarisuusarvot viittaavat, että graafissa on löydettävissä yhteisörakenne. [Fortunato and Castellano, 2007.]

Papadopouloksen ja muiden [2011] mukaan lokaali näkökulma saattaa olla sosiaalisen median yhteisöjen näkökulmasta globaalia merkityksellisempi, koska tutkittavasta verkostosta on usein vain osittainen tietämys.

4.3. Muita yhteisöjen määrittelytapoja

Yhteisöt voidaan määritellä myös solmujen samankaltaisuuden perusteella. Tällöin jokainen solmu päätyy sellaiseen klusteriin, jonka muut solmut ovat eniten samankaltaisia solmun kanssa. Jokaisen solmuparin samankaltaisuus voidaan laskea jonkin lokaalin tai globaalin ominaisuuden perusteella, esimerkiksi solmujen välisiä etäisyyksiä voidaan arvioida. [Fortunato and Castellano, 2007.] Samankaltaisuuden mittaaminen ei edellytä, että solmuparin välillä on

särmä. Perinteiset klusterointimenetelmät, kuten hierarkkinen ja osittava käyttävät perustanaan samankaltaisuusmittoja. [Fortunato, 2010.] Klusterointimenetelmiin perehdytään tarkemmin seuraavassa luvussa.

Yhteisöt voidaan määritellä myös tarkastelemalla niiden muodostumisprosessia verkostoissa [Papadopoulos *et al.*, 2011]. Esimerkiksi Pallan ja muiden [2005] esittelemä *klikkien perkolaatiomenetelmä* (clique percolation method) määrittelee yhteisöt toisiinsa yhteydessä olevien *k-klikkien* (*k*-clique) unioneina. Tällöin *k*-klikki tarkoittaa täydellistä, maksimaalista aligraafia, jossa on *k* solmua. Toisiinsa yhteydessä olevien *k*-klikkien tulee jakaa *k* - 1 solmua. [Palla *et al.*, 2005.] Tämä menetelmä on suosituin limittäisten yhteisöjen löytämiseen käytetty tekniikka [Fortunato, 2010].

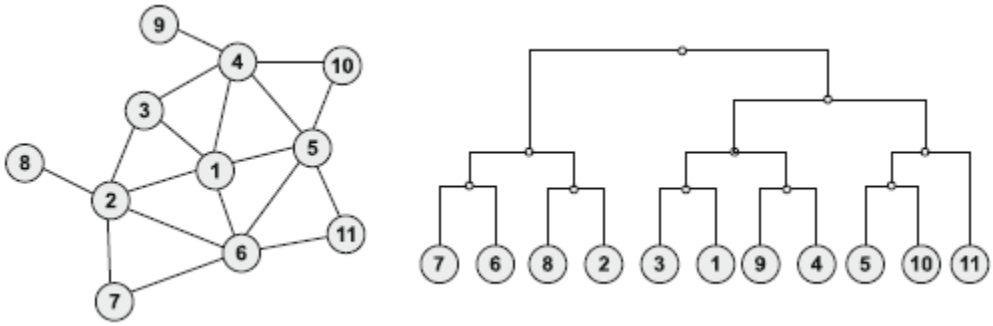
Yhteisöjä voidaan siis määritellä monella tavalla, mutta monissa yhteisöjen löytämisalgoritmeissa yhteisöille ei ole tarkkaa ennakkomääritelmää. Tällöin yhteisöt ovat vasta menetelmästä syntyneitä lopputuotteita [Fortunato, 2010].

5. Graafien klusterointimenetelmiä

Graafin klusteroinnilla tarkoitetaan graafin solmujen ryhmittelyä sen rakenteen mukaisesti [Webb and Copsey, 2011]. Sosiaalisen median dataan soveltuvien klusterointimenetelmien löytäminen on haastavaa, koska monissa menetelmissä edellytetään klustereiden lukumäärän antamista syötteenä. Papadopouloksen ja muiden [2011] mielestä yhteisöjen lukumäärää on kuitenkin lähes mahdotonta arvioida etukäteen sosiaalisen median laajoja ja dynaamisia verkostoja klusteroitaessa, vaan yhteisöjen lukumäärän tulee olla pikemminkin menetelmän lopputulos. Sosiaalisten verkostojen yhteisöjen löytämisessä on käytetty perinteisesti osittavaa ja hierarkkista klusterointimenetelmää, joissa solmut ryhmitellään niiden keskinäisen samankaltaisuuden perusteella.

5.1. Osittava klusterointi

Graafien osittamisessa (graph partitioning) graafi jaetaan osiin eli klustereihin siten, että jokainen solmu kuuluu yhteen klusteriin. Kuitenkin todellisuudessa solmut voivat limittäisten yhteisöjen tapauksissa kuulua useampaan eri ryhmään. Tällaista graafin jakoa limittyviin yhteisöihin kutsutaan *peitoksi* (cover). Ositetut graafit voidaan järjestää hierarkkisesti, jolloin rakenne koostuu monista sisäkkäisistä yhteisöistä. Graafin hierarkkista rakennetta voidaan esittää *dendrogrammin* (dendrogram) avulla, josta on esimerkki kuvassa 7. [Fortunato, 2010.] Dendrogrammi on puu, jonka lehdet kuvaavat solmuja ja oksat niitä yhdistäviä särmiä. Korkeammalla tasolla oksat yhdistävät solmujen joukkoja havainnollistaen yhteisöjen hierarkkista rakennetta ja sisäkkäisiä yhteisöjä. [Radicchi *et al.*, 2004.] Radicchi ja muut [2004] kuitenkin huomauttavat, että puut eivät sinällään kuvaa, mitkä oksista ovat todella merkittäviä, vaan verkostosta tarvitaan lisätietoa yhteisöjaon luotettavuuden arvioimiseksi.



Kuva 7. Yksinkertainen graafi ja rakennetta vastaava dendrogrammi [Papadopoulos *et al.*, 2011].

Graafien jakaminen osiin on haastava tehtävä, koska algoritmien tulee löytää mahdollisimman hyvät ositukset lukuisista eri vaihtoehdoista. Mahdollisten graafin ositusten määrä nimittäin kasvaa nopeammin kuin eksponentiaalisesti graafin kokoon nähden. Yhteisöjen mahdollinen hierarkkisuus ja osittainen limittäisyys tuovat omat haasteensa yhteisöjen löytämiselle. [Fortunato and Castellano, 2007.]

Duanin ja muiden [2009] mukaan graafin hyvässä ja tehokkaassa osituksessa on tärkeää valita sopiva mittari yhteisön evaluointia varten, joka ottaa huomioon yhteisön yleisen rakenteen ja käyttää aikaa säästäviä algoritmeja optimaalisen osituksen löytämiseksi. Osituksen hyvyyttä voidaan evaluoida vaikkapa modulaarisuuden mittarilla. Optimaalinen modulaarisuus on Duanin ja muiden [2009] mukaan välillä 0,3–0,7. Paras graafin ositus on sellainen, jonka modulaarisuus on suurin. Kuitenkin jo alaraja 0,3 osoittaa, että yhteisörakenne on huomionarvoinen [Duan *et al.*, 2009].

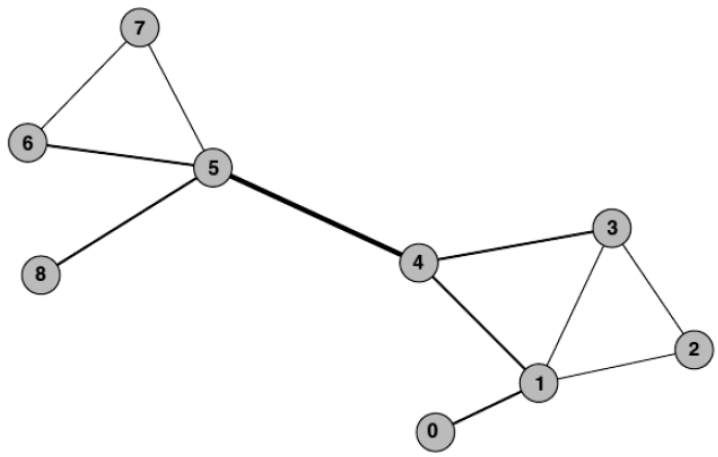
5.2. Hierarkkinen klusterointi

Hierarkkiset klusterointimenetelmät perustuvat yleensä objektiparien samankaltaisuus- tai erilaisuusmatriiseihin tehtävään analysointiin. Samankaltaisuusmittarit voivat olla rakenteellisia, jolloin huomioidaan solmujen välisten yhteyksien muodostamat mallit, tai attribuuttiperusteisia, jolloin käytetään solmujen ominaisuuksia. [Webb and Copsey, 2011.] Hierarkkinen klusterointi voidaan erotella *jakaviin* (divisive) ja *kokoaviin* (agglomerative) menetelmiin.

Jakavissa klusterointimenetelmissä kokonainen verkosto jaetaan yhteisöihin poistamalla särmiä yksi kerrallaan. Tällöin verkosto jakautuu progressiivisesti

yhä pienempiin toisistaan erillisiin yhteisöihin [Radicchi *et al.*, 2004.] Suurille verkostoille nämä menetelmät saattavat olla laskennallisesti raskaita ja kalliita [Webb and Copsey, 2011]. Merkittävin tekijä jakavassa klusteroinnissa on irrotettavien särmien valinta, koska niiden tulee olla yhteisön ulkopuolisia, mutta yhteisöjä yhdistäviä. Eräs jakavista lähestymistavoista on Girvanin ja Newmanin [2002] kehittämä menetelmä, josta on Radicchin ja muiden [2004] tutkimusten mukaan muodostunut eräänlainen standardi yhteisöjen rakenteen analysoinnissa.

Girvanin ja Newmanin kehittämä algoritmi jakaa verkostot yhteisöihin tunnistamalla yhteisöjen väliset särmät ja poistamalla ne iteratiivisesti. Kyseisten särmien tunnistamisessa käytetään mittarina särmien välillisyysarvoa [Fortunato, 2010.] *Särmän välillisyys* (edge betweenness) kuvaa särmää pitkin kulkevien solmuparien välisten geodeesisten polkujen määrää. Jos solmuparin välillä on useampi geodeesinen polku, niille annetaan sama painoarvo. [Girvan and Newman, 2002.] Perustana tässä algoritmissa on lähtökohta, että eri yhteisöt verkostossa ovat toisissaan kiinni heikosti vain ehkä muutaman särmän avulla, jolloin myös lyhimmät polut eri yhteisöjen solmujen välillä kulkevat näitä särmiä pitkin. Täten kyseisillä särmillä on korkea välillisyysarvo, jonka perusteella ne voidaan paikallistaa ja poistaa. [Webb and Copsey, 2011.] Menetelmä perustuu siis sellaisten särmien löytämiseen, jotka sijaitsevat yhteisöjen välillä [Girvan and Newman, 2002]. Välillisyyksiä on havainnollistettu kuvassa 8, jossa särmän paksuus on suhteessa särmän välillisyysarvon suuruuteen. Paksuin särmä solmujen 4 ja 5 välillä yhdistää verkoston kahta eri yhteisöä ja sillä on korkein välillisyysarvo, koska se on kaikkien kahden yhteisön välillä kulkevien polkujen varrella.



Kuva 8. Kahden yhteisön verkosto [Webb and Copsey, 2011].

Girvanin ja Newmanin jakavassa algoritmissa särmät, joilla on suurin välillisyysarvo, poistetaan siis ensin. Poistaminen jakaa verkoston irrallisiin aligraafeihin, joissa suoritetaan sitten sama prosessi. Särmän poistamisen jälkeen välillisyysarvot tulee laskea uudelleen kaikille niille särmille, joihin edeltävä särmän poistaminen vaikuttaa. Menetelmä on tämän vuoksi melko raskas, ja soveltuu siten parhaiten suhteellisen pienille graafeille. [Girvan and Newman, 2002.] Iteraatioita jatketaan, kunnes koko verkosto on jaettu irrallisten solmujen joukkoon. Dendrogrammi rakennetaan jakavassa menetelmässä siten juuresta lehtiin päin. [Radicchi *et al.*, 2004.]

Kokoavissa klusterointimenetelmissä solmuja ryhmitellään isommiksi yhteisöiksi, kunnes koko verkosto on rakennettu. Kokoavassa menetelmässä jokaiselle verkoston solmuparille lasketaan paino, joka mittaa solmujen välistä läheisyyttä. Klusterointi aloitetaan kaikkien solmujen joukosta ja solmuparien välille lisätään linkkejä iteratiivisesti suurimmasta painosta alkaen laskevassa järjestyksessä. Dendrogrammi rakennetaan lehdistä juureen päin, joka edustaa siten koko verkostoa. [Radicchi *et al.*, 2004.]

Klusterointimenetelmiä voidaan käyttää datan jakautumisen tarkasteluun ja analysointiin sellaisenaan, jolloin verkostoista muodostuneet yhteisöt ovat menetelmien tavoitteita. Klusterointi ja yhteisöjen löytäminen saattavat kuitenkin olla vain välivaihe ja datan esikäsittelyä, jolloin muodostuneita klustereita halutaan analysoida yksityiskohtaisemmin ja pienemmissä osissa. Seuraavaksi perehdytään yhteisöjen analysointiin.

6. Yhteisöjen analysointi

Suurin osa verkostoista on dynaamisia [Webb and Copsey, 2011]. Verkostojen erilaiset yhteisöt syntyvät, muuttuvat ja häviävät. Yhteisöjä ja niiden dynaamisuutta voidaan tutkia linkkianalyysin avulla.

6.1. Linkkianalyysi

Linkkien louhinta ja analysointi sisältävät kuvailevan ja ennustavan mallinnuksen, jossa objektien välisiä linkkejä eli suhteita tutkitaan erilaisten menetelmien avulla. Han ja Kamber [2006] esittelevät erilaisia linkkien louhinnan tehtäviä:

1. *Linkkipohjainen objektien luokittelu*. Perinteisissä menetelmissä objektit luokitellaan jonkin attribuutin perusteella, mutta linkkipohjaisessa menetelmässä käytetään attribuuttien lisäksi linkkejä objektin luokan ennustamiseen. Esimerkiksi verkkosivu voidaan luokitella sen sanojen esiintymien ja ankkuritekstien lisäksi sivujen välisten linkkien perusteella.

2. *Objektityypin ennustaminen*. Objektin tyyppi voidaan ennustaa sille ominaisten attribuuttien lisäksi siihen linkeillä yhdistettyjen muiden objektien attribuuttien avulla. Esimerkiksi kontaktitiedosta voidaan ennustaa, onko se puhelin-, sähköposti- vai osoiteyhteystieto.

3. *Linkkityypin ennustaminen*. Linkin tyypin tai tarkoituksen ennustamista varten tutkitaan siihen liittyvien objektien ominaisuuksia. Verkkosivujen linkeistä voidaan esimerkiksi ennustaa mainos- ja navigointilinkit. Myös toisensa tuntevista ihmisistä voidaan pyrkiä päättelemään, ovatko he perheenjäseniä, tuttavia tai työtovereita. Linkkityypin ennustamista varten objektien välinen yhteys on havaittu.

4. *Linkkien ennustaminen*. Linkkien ennustamisessa pyritään selvittämään, löytyykö objektien väliltä linkki. Voidaan vaikkapa ennustaa, onko kahden verkkosivun välillä yhteys. Linkkien ennustamista selvitetään lisää yhteisöjen evoluution yhteydessä kohdassa 6.2.

5. *Linkin kardinaliteetin arvioiminen.* Linkin kardinaliteetti voi tarkoittaa kahta eri lukumäärää. Voidaan arvioida tiettyyn objektiin osoittavien

linkkien määrä, jonka avulla pystytään ennustamaan esimerkiksi verkkosivun vaikutusvalta. Toisaalta objektista lähtevien linkkien määrän avulla voidaan paikallistaa *keskipisteinä* (hubs) toimivat objektit. Vaikeammin arvioitava kardinaliteetti on objektien lukumäärä, jotka osuvat tietystä objektista lähtevän linkin polulle. Tällainen polku voi muodostua esimerkiksi verkkosivuista, jotka yhdistyvät toisiinsa linkkipolun avulla.

6. *Objektin tunnistaminen*. Objektin tunnistamisella pyritään selvittämään, ovatko kaksi objektia itse asiassa sama entiteetti niiden attribuuttien ja linkkien perusteella. Tätä menetelmää voidaan käyttää esimerkiksi duplikaattien poistamisessa.

7. *Ryhmän tunnistaminen*. Ryhmän tunnistamista käytetään klusteroinnissa, jossa joukko objekteja arvioidaan kuuluvan samaan klusteriin niiden attribuuttien ja linkkirakenteen perusteella. Tehtävää sovelletaan esimerkiksi tiettyyn teemaan tai aiheeseen keskittyvien verkkoyhteisöjen löytämisessä.

8. *Aligraafin tunnistaminen*. Aligraafin tunnistaminen on merkittävien aligraafien löytämistä verkostoista. Tunnistamisen perusteena voidaan käyttää vaikkapa tietynlaista rakennetta graafissa. Yhteisöjen löytäminen on itse asiassa aligraafin tunnistamista.

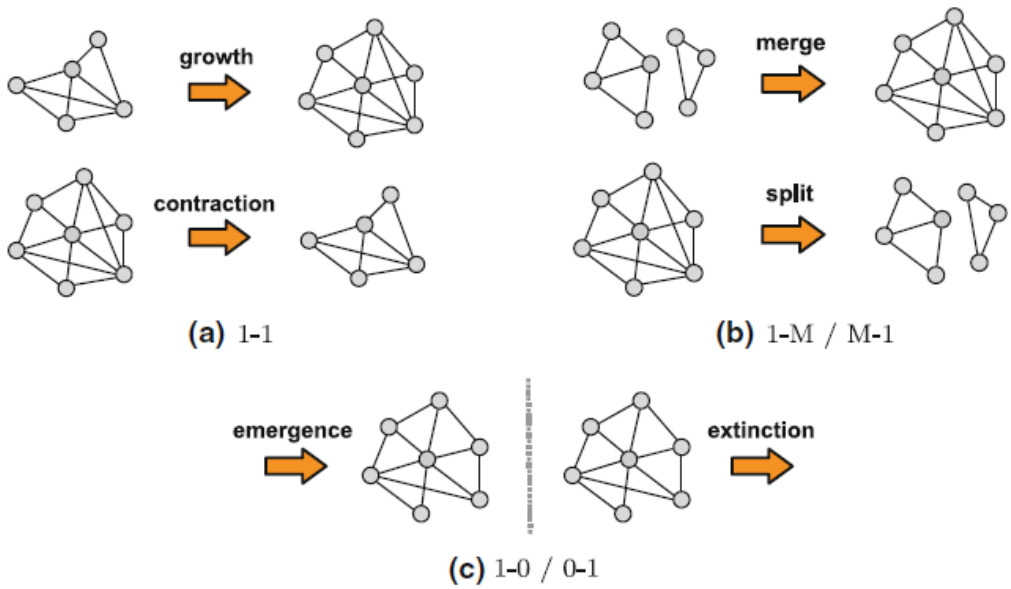
9. *Metadatan louhinta.* Puolirakenteellista metadataa voidaan hyödyntää datan integraatiotehtävissä monilla eri sovellusalueilla. Metadatan avulla voidaan kartoittaa kahden eri lähteen dataa ja löytää niistä dataa, joka kuuluu samalle objektille.

6.2. Yhteisöjen evoluutio

Varsinkin sosiaalisen median verkostoihin liittyvä dynaamisuus ja ajallinen evoluutio edellyttävät muutakin kuin yhteisöjen staattisen rakenteen analysointia. Kuvassa 9 on esitelty yhteisön perustoiminnot, jotka sisältyvät Papadopouloksen ja muiden [2011] mukaan yhteisön evoluutioon. Kuvassa näkyy myös heidän esittelemänsä kolme eri muutostyyppiä: (a) yhdestä yhdeksi (one-to-one), jossa yhteisö joko kasvaa tai kehittyy, (b) yhdestä moneksi (one-to-many), jossa yhteisö voi joko jakautua useampaan tai useampi yhteisö voi yh-

distyä yhdeksi (monesta yhdeksi, many-to-one) ja (c) yhdestä nollaksi (one-to-zero), jossa yhteisö joko häviää tai syntyy (nollasta yhdeksi, zero-to-one). Evoluution kuusi perustoimintoa ovat:

1. yhteisön kasvaminen (growth)
2. yhteisön selviytyminen (contraction/survive)
3. usean yhteisön yhdistyminen (merge)
4. yhteisön jakautuminen useampaan yhteisöön (split)
5. uuden yhteisön muodostuminen (emergence/form)
6. yhteisön häviäminen (extinction/dissolve)

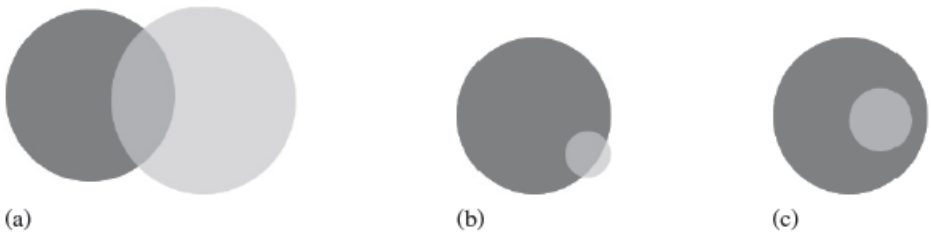


Kuva 9. Yhteisön evoluution perustoiminnot [Papadopoulos *et al.*, 2011].

Yhteisöjen evoluutiota voidaan analysoida linkkien ennustamisen avulla. Tällöin yhteisöstä otetaan tilannekuvaus tiettynä ajankohtana ja pyritään arvioimaan, mitä särmiä verkostoon tulee lisää määrätyn ajanjakson kuluessa. Linkkejä voidaan pyrkiä ennustamaan esimerkiksi sosiaalisten yhteisöjen ystävyyssuhteiden ennakoimisessa [Webb and Copsey, 2011]. Esimerkiksi Facebookissa voidaan ennustaa, kuinka todennäköisesti kahdesta käyttäjästä tulee ystäviä. Hanin ja Kamberin [2006] mukaan linkkien ennustamismenetelmissä käytetään *yhteyspainoa*(v_i, v_j) solmuille v_i ja v_j perustuen valittuun läheisyysmittariin ja tutkittavaan graafiin. Yhteyspainoista muodostetaan sitten järjestetty lista, jonka perusteella uusia linkkejä voidaan ennustaa. Kokeellisia datajoukkoja havainnoimalla voidaan sitten evaluoida ennustusten oikeellisuutta. Läheisyysmitta-

rina voidaan käyttää esimerkiksi *lyhintä polkua* (shortest path) solmuparin välillä tai *yhteisten naapureiden* (common neighbors) lukumäärää. Mitä enemmän solmuilla v_i ja v_j on yhteisiä naapureita, sitä todennäköisemmin niiden välillä tulee joskus olemaan linkki. Muut mittarit ottavat huomioon Hanin ja Kamberin [2006] mukaan kahden solmun välisten polkujen kokonaisuuden, esimerkiksi poluista voidaan laskea painotettu summa. Kaikkia edellä mainittuja mittareita voidaan myös käyttää yhdessä klusteroinnin kanssa.

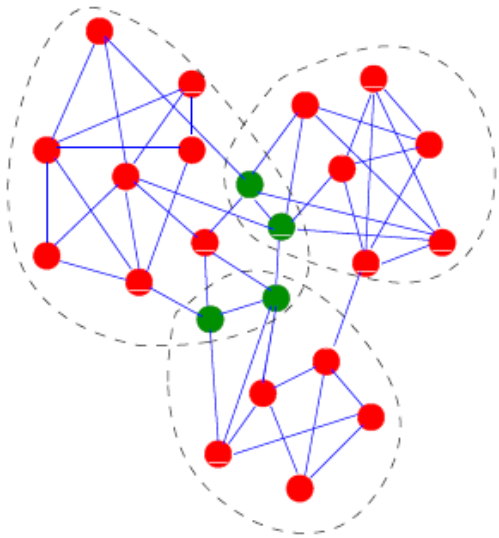
Linkkien ennustamisen lisäksi yhteisöjen tilannekuvauksia voidaan tarkastella sekvensseinä, joiden perusteella yhteisöjen perustoimintoja pystytään havainnoimaan. Tällaisessa ajallisesti järjestetyssä sekvenssissä on tilannekuvauksia yhteisön syntymisen ajanjaksosta siihen hetkeen, kun yhteisö on viimeksi havaittu. Takaffoli ja muut [2011] käyttävät tästä myös nimitystä *metayhteisö* (meta community). Tärkein käsite tilannekuvauksien analysoinnissa on yhteisöjen *samankaltaisuus* (similarity). Kahden eri tilannekuvauksen yhteisöt ovat samankaltaisia, jos niiden jäsenistä tietty määrä on samoja. Tämä prosentuaalinen kynnysarvo voidaan asettaa tarkasteltavan verkoston ominaisuuksien perusteella. Samalla kynnysarvo määrää, kuinka paljon yhteisön jäsenistö voi vaihdella. Vakaissa yhteisöissä on paljon pitkäkestoisia jäsenyyksiä ja vähän vaihtuvuutta, jolloin samankaltaisuuskynnysarvon voi olettaa olevan korkea. Sen sijaan hyvin dynaamisissa verkostoissa on havaittavissa epävakaita yhteisöjä, joissa vanhoja jäseniä lähtee asteittain samalla, kun uusia jäseniä liittyy. Tällainenkin yhteisö voi säilyä pitkän aikaa, vaikka alkuperäisiä jäseniä ei enää olisi yhtään. Täten epävakaiden yhteisöjen samankaltaisuuskynnysarvon olisi hyvä olla matala. [Takaffoli *et al.*, 2011.] Kuvassa 10 on havainnollistettu yhteisöesimerkkejä samankaltaisuuden mittaamiseksi: a) Yhteisöissä 110 ja 120 jäsentä, joista 30 yhteistä jäsentä; b) Yhteisöissä 100 ja 30 jäsentä, joista 20 yhteistä jäsentä; c) Yhteisöissä 100 ja 40 jäsentä, joista 40 yhteistä jäsentä.



Kuva 10. Yhteisöpareja samankaltaisuuden mittaamiseksi [Takaffoli *et al.*, 2011].

Sosiaaliset verkostot jakautuvat yleensä hierarkkisesti yhteisöihin, jotka puolestaan sisältävät pienempiä yhteisöjä jne. Papadopoulos ja muut [2011] mainitsevat, että hierarkkisia rakenteita löytää varsinkin sosiaalisen median sovelluksista. Esimerkkinä he mainitsevat yhteisön, joka koostuu jonkin tietyn indie rock -yhtyeen faneista, jotka samanaikaisesti voidaan katsoa kuuluvan laajempaan rockmusiikki-fanien yhteisöön.

Sosiaalisten verkostojen käyttäjät kuuluvat tyypillisesti useampiin eri yhteisöihin samanaikaisesti, esimerkiksi työhön, harrastuksiin ja ystävyyssuhteisiin liittyen. Tällaisesta ominaisuudesta käytetään nimitystä *yhteisöjen limittäisyys* (overlapping communities). Yksittäistä solmua ei tällöin voida merkitä vain yhteen ryhmään kuuluvaksi, jottei mahdollisesti relevanttia informaatiota hukattaisi [Fortunato, 2010]. Kuvassa 11 on esimerkki kolmesta yhteisöstä, jotka limittyvät toisiinsa siten, että neljä solmua kuuluu kahteen eri yhteisöön. Chen ja muut [2009] jakavat tyypillisten sosiaalisten verkostojen solmut kolmeen eri kategoriaan: yhteisöihin, keskipisteisiin ja *erakoihin* (outliers). Keskipisteet ovat solmuja, jotka kuuluvat useampaan eri yhteisöön ja yhdistävät siten ryhmiä toisiinsa verkostossa. Chen ja muut [2009] määrittelevät erakot solmuiksi, jotka eivät kuulu mihinkään yhteisöön. Heidän mukaansa kyseisten kategorioiden tunnistaminen on oleellista yhteisöjen tunnistamiseen liittyvissä sovelluksissa.



Kuva 11. Limittyvät yhteisöt [Fortunato and Castellano, 2007].

6.3. Yhteisöjen ominaisuuksia ja mittareita

Sosiaalisia verkostoja ja yhteisöjä voidaan analysoida erilaisten mittareiden ja ominaisuuksien avulla. Kvantitatiivisten mittareiden avulla voidaan vertailla verkostoja, tutkia verkostojen rakennetta ja havaita ajallisia muutoksia verkostoissa [Hansen *et al.*, 2011]. Seuraavaksi esitellään merkittävimpiä yhteisöjen attribuutteja ja mittareita:

- *Solmujen yhtenäisyysaste* (vertex connectivity) on pienin määrä solmuja, joiden poistaminen tekee graafista epäyhtenäisen tai yksisolmuisen [Koivisto ja Niemistö, 2001]. Yhtenäisyysaste on *yhteisön sisällä* (intercommunity) korkea ja *yhteisöjen välillä* (between-community) matala. Tämä kuvastaa yhteisörakennetta. [Webb and Copsey, 2011.] Yhteisön yhtenäisyydestä voidaan myös käyttää nimitystä *koheesio* (cohesion) tai *rakenteellinen yhtenäisyys* (structural cohesion) [Hansen *et al.*, 2011]. Korkea koheesio tarkoittaa sitä, että yhteisöstä pitäisi poistaa lukuisia särmiä, ennen kuin yhteisö jakautuisi erillisiin komponentteihin [Yang and Leskovec, 2012].

- Yhteisön *tiheys* (density) kuvaa, kuinka paljon solmujen välillä on särmiä verrattuna siihen, paljonko niitä yhteisöä vastaavassa aligraafissa enimmillään voisi olla. Yksinkertaisen, suuntaamattoman graafin tiheys voidaan laskea kaavalla

$$\frac{m}{n(n-1)/2},$$

  jossa m on graafissa olevien särmien ja n solmujen lukumäärä. Solmuparia v_i ja v_j yhdistävä särmä on sama kuin solmuparia v_j ja v_i yhdistävä särmä. Graafin maksimaalinen särmien lukumäärä voidaan siis laskea kaavalla $n(n-1)/2$. Tiheys vaihtelee välillä 0–1, jolloin täydellisen graafin tiheys on 1. [Scott, 2000]. Yksinkertaisen, suunnatun graafin tiheys lasketaan kaavalla

$$\frac{m}{n(n-1)},$$

  koska särmien suurin mahdollinen määrä on sama kuin eri solmuparien lukumäärä [Scott, 2000].

- Yhteisön *sentralisaatio* (centralization) tarkoittaa ominaisuutta, jossa verkosto on keskittynyt vain yhden tai muutaman tärkeän solmun ympärille, jolloin näihin liittyy myös paljon särmiä. Sen sijaan *desentralisoiduissa* (decentralized) yhteisöissä on vain vähän vaihtelua solmuista kulkeutuvien särmien määrissä. [Hansen *et al.*, 2011.]

- Yhteisöjä voidaan arvioida monilla muillakin mittareilla, esimerkiksi *verkoston sietokyky* (network resilience) kuvaa, miten hyvin verkosto toipuu solmujen tai särmien poistamisesta.

- Suunnatusta graafista voidaan tutkia myös yhteisön *vastavuoroisuutta* (reciprocity) eli solmujen välisten yhteyksien molemminpuolisuutta.

Verkostoa voidaan kuvata myös *ydin-reuna-alue -rakenteen* (core-periphery structure) avulla. Csermelyn ja muiden [2013] mukaan verkoston ytimellä tarkoitetaan keskeistä ja toistensa kanssa tiiviisti yhteydessä olevaa solmujen joukkoa. He käyttävät sosiaalisten yhteisöjen ytimestä nimitystä ”eliitti”. Eliitin jäsenet ovat myös usein niitä, jotka muodostavat uusia ryhmiä. Yhteisön uudet jäsenet ovat usein yhteydessä muiden reuna-alueelle sijoittuvien jäsenten kanssa, mutta muodostavat yhteyksiä eliittiin myöhemmin. Chakrabartin [2003] mukaan suuret ytimet kuvaavat jo vakiintuneita yhteisöjä. Sen sijaan syntyvät, pienet ydinalueet saattavat kuvata hänen mukaansa täysin uudenlaisia yhteisöjä. Kaikki keskeiset solmut eivät kuitenkaan muodosta verkoston ydintä. Esimerkiksi kahden yhteisön välillä sijaitseva solmu voi olla keskeinen, koska se yhdistää yhteisöjä toisiinsa. Yhteisön reuna-alueilla sijaitsevat solmut saattavat olla vain yhdellä särmällä kiinni verkostossa. Solmujen erilaisista rooleista yhteisöissä kerrotaan tarkemmin kohdassa 7.1.

Myös yhteisöön liittyvää attribuuttidataa voidaan analysoida. Esimerkiksi solmujen *homofiliaa* (homophily) voidaan mitata, jolla pyritään selvittämään yhteydessä olevien yksilöiden samankaltaisuus. Hansenin ja muiden [2011] mukaan tutkimukset osoittavat, että tyypillisesti ihmiset ovat yhteydessä perusominaisuuksiltaan samankaltaisten henkilöiden kanssa. Näihin ominaisuuksiin kuuluvat heidän mukaansa esimerkiksi ikäryhmä, uskonto, tulo- ja koulutustaso.

7. Solmujen ja särmien analysointi

Yhteisöjen ja niiden rajojen löytäminen mahdollistaa pienempien alirakenteiden analysoimisen. Koko yhteisön ominaisuuksien lisäksi yhteisöstä saatetaan etsiä vain tietynlaisia mielenkiinnon kohteena olevia solmuja tai särmiä. Solmuja voidaan esimerkiksi luokitella niiden rakenteellisen sijainnin perusteella, jolloin solmuille määritellään erilaisia rooleja. Solmuja voidaan analysoida myös tutkimalla niiden ”käyttäytymistä”, eli miten solmut muodostavat yhteyksiä muihin solmuihin.

Poikkeavat solmut (anomalous nodes) saattavat erota muista solmuista epätavallisen korkean asteen tai poikkeavalla tavalla muodostuvien yhteyksien vuoksi. Lisäksi solmu saattaa väliaikaisesti käyttäytyä erikoisella tavalla, joka poikkeaa aiemmasta käytöksestä. Esimerkiksi yhteyksien määrä vierussolmuihin saattaa muuttua poikkeavalla tavalla. [Webb and Copsey, 2011.]

Solmujen ja eri yhteisöjen välillä kulkevista särmistä voidaan löytää *merkittäviä* (significant) ja *poikkeavia särmiä* (anomalous edges). Merkittävät särmät yhdistävät verkoston toiminnan kannalta olennaisia solmuja, ja poikkeavat särmät ovat jollain tavalla epätavallisia verkoston toiminnassa. Lisäksi joidenkin solmujen välille oletettu särmä saattaa kokonaan puuttua. [Webb and Copsey, 2011.]

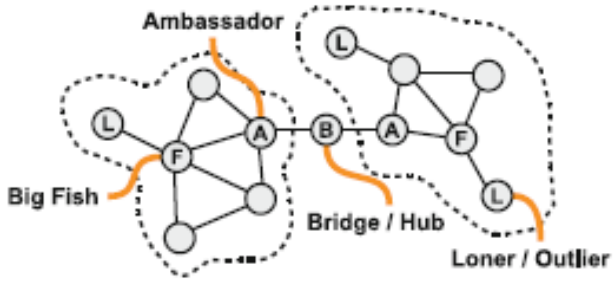
7.1. Solmujen roolit yhteisöissä

Yksilöä edustavan solmun sijainti suhteessa muihin solmuihin on Hansenin ja muiden [2011] mukaan sosiaalisten verkostojen analysoinnin ensisijainen mielenkiinnon kohde. Heidän mielestään analysoinnin tavoitteena on usein yksilön yhteysmallien hahmottaminen sen sijaan, että keskityttäisiin yksilön ominaispiirteisiin. Yhteysmallit kuvaavat sitä, kuinka yksilöt usein toimivat samalla tavalla samankaltaisissa olosuhteissa ja sosiaalisissa tilanteissa [Hansen *et al.*, 2011]. Klusterin keskellä sijaitsevat solmut yhdistyvät useiden muiden solmujen kanssa, joten kyseisillä solmuilla saattaa olla yhteisössä tärkeä rooli ryhmän tasapainottajana ja vakauttajana. Klusterin laidoilla sijaitsevat solmut puolestaan voivat toimia merkittävässä roolissa eri yhteisöjen välisessä kanssakäymisessä. [Fortunato, 2010.] Solmun roolilla tarkoitetaan siis kuvausta siitä, millaisessa osassa solmu on verkostorakenteessa, eli miten solmu käyttäytyy suhteessa sen naapureihin ja laajempaan verkostoon. Tällaista tietoa voidaan hyödyn-

tää *vaikutusvallan maksimoimisessa* (influence maximization) ja linkkipohjaisissa luokitteluissa esimerkiksi internethakuihin, sitaattianalyyseihin ja rikostutkintaan sekä terrorismiuhkien havaitsemiseen liittyen.

Vaikutusvallan maksimoinnilla tarkoitetaan keskeisten ”vaikuttajayksilöiden” löytämistä, josta voi olla hyötyä esimerkiksi tuotteen markkinoinnissa tai uuden idean läpiviennissä [Scripps *et al.*, 2007]. Yhteisön jäsenten käyttäytymistä toisiinsa vaikuttamisen näkökulmasta voidaan mallintaa suunnatulla vaikutusvaltagraafilla, jossa jokaista ryhmän jäsentä edustaa yksi solmu. Kaari merkitään solmusta v_i solmuun v_j silloin, kun solmua v_i edustava henkilö vaikuttaa solmua v_j edustavan henkilön käyttäytymiseen [Koivisto ja Niemistö, 2001]. Vaikutusvaltaa kuvaavat graafit ovat siis usein suunnattuja ja särmiä on painotettu. Täten solmun vaikutusvaltaa, mainetta tai statusta kuvataan ensisijaisesti sen lähtö- ja tuloasteella. [Chakrabarti, 2003.] *Vaikutusvaltaisia solmuja* (influential nodes) voidaan löytää myös suuntaamattomista graafeista solmujen sijainnin ja yhteyksien perusteella. Tällaiset solmut ovat verkostojen toiminnan kannalta keskeisiä ja niiden poistamisella on merkittävää vaikutusta verkoston käyttäytymiseen. Vaikutusvaltainen solmu saattaa esimerkiksi yhdistää kahta yhteisöä ja toimia niiden välisenä yhteydenpitäjänä. Tällaisen solmun poistaminen jakaisi yhteisöt kokonaan erilleen. [Webb and Copsey, 2011.] Toisaalta kommunikointia tai informaation kulkua kuvaavissa verkostoissa vaikutusvaltainen solmu saattaa omasta tahdostaan esimerkiksi vääristellä tai pitää tietoa itsellään, koska se pystyy kontrolloimaan yhteydenpitoa [Freeman, 1979].

Scripps ja muut [2007] esittelevät neljä erilaista yhteisöissä esiintyvää roolia: *silta/keskipiste* (bridge/hub), *puolestapuhuja* (ambassador), *iso kiho* (big fish) ja erakko (loner/outlier), joita havainnollistetaan kuvassa 12. Keskipisteillä ei ole paljon yhteyksiä muiden solmujen kanssa, mutta ne toimivat siltoina eri yhteisöjen välillä. Puolestapuhujilla on yhteyksiä useiden eri yhteisöjen solmuihin, mutta isoilla kihoilla on yhteyksiä vain useiden saman yhteisön solmujen kanssa. Erakoilla on nimensä mukaisesti vain vähän yhteyksiä eri solmuihin ja yhteisöihin. Joskus verkoston syrjässä olevat solmut eivät kuulu mihinkään yhteisöön [Chen *et al.*, 2009].



Kuva 12. Yhteisöissä esiintyviä rooleja [Papadopoulos *et al.*, 2011].

Webbin ja Copseyn [2011] mukaan *keskeiset solmut* (central nodes) ovat yhteydessä monien muiden solmujen kanssa. He määrittelevät myös keskeisten solmujen vastakohdiksi syrjässä olevat solmut, joilla on vain vähän yhteyksiä muihin solmuihin. Tulee kuitenkin huomioida, että solmu voi olla keskeinen, vaikka sillä ei olisi paljon yhteyksiä muihin solmuihin. Esimerkiksi kuvassa 12 esiintyvä solmu B eli kahden yhteisön välissä toimiva silta on keskeinen solmu. Solmujen keskeisyyden eri näkökulmia ja mittareita kuvataan tarkemmin kohdassa 7.2.

Solmun sijainnin lisäksi sen merkittävyyttä yhteisöissä voidaan arvioida painotetun vuorovaikutuksen mukaan. Tällöin solmujen välisille särmille lasketaan painotettu arvo jonkin ominaisuuden mukaan, esimerkiksi kahden ihmisen välistä viestienvaihdon määrää voidaan kuvata painon avulla. Tällainen painoarvo on temporaalinen ja muuttuu esimerkiksi ystävyyssuhteen kehittyessä [Webb and Copsey, 2011].

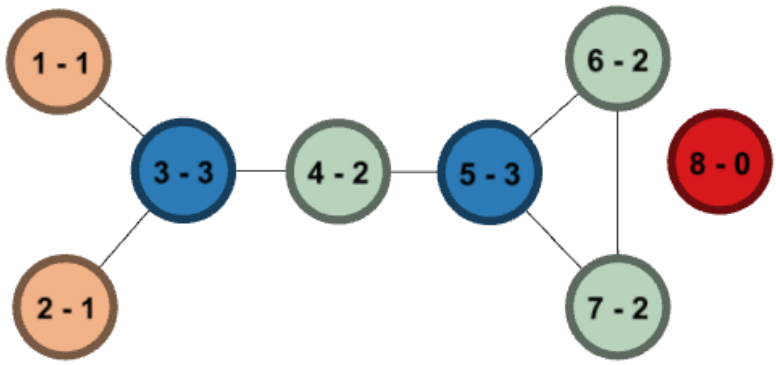
Seuraavaksi esitellään mittareita, joiden avulla voidaan määritellä yksittäisten solmujen positioiden perusteella yksilöiden rooleja ja vaikutusvaltaa yhteisöissä [Hansen *et al.*, 2011].

7.2. Solmukohtaiset keskeisyysmittarit

Solmun *keskeisyys* (centrality) kuvaa solmun tärkeyden tai osallistumisen määrää verkostossa. Keskeisyyttä kuvaavia mittareita ja niiden muunnoksia on lukuisia, eikä niitä kaikkia siten voida tässä kuvata. Sosiaalisten verkostojen solmujen keskeisyyden mittareina käytetään yleisimmin astetta, läheisyyttä, välil-

lisyyttä ja ominaisvektoria. Keskeisyyttä kuvaavan mittarin valinta riippuu verkoston tyypistä ja sovellusalueesta sekä tutkimuskohteista.

Aste on yleensä se ensimmäinen (ja joskus ainoa) mittari, jonka avulla solmun keskeisyyttä arvioidaan. Solmuun liittyvien särmien lukumäärä voidaan laskea, vaikka tunnettaisiin vain solmun välitön ympäristö verkostossa [Opsahl *et al.*, 2010]. Kuvassa 13 on yksinkertainen, epäyhtenäinen graafi, jonka solmuille on laskettu asteet. Solmuun merkitty ensimmäinen numero on solmun ID-tunnus ja toinen numero solmun aste. Solmuilla 3 ja 5 on graafissa suurin aste: 3.



Kuva 13. Yksinkertainen, epäyhtenäinen graafi, johon on merkitty solmujen asteet.

Aste on yleensä helppo intuitiivisesti ymmärtää, ja siten tutkijat käyttävät sitä usein selittäessään keskeisyyttä verkostometriikkaa vähemmän tuntevalle kohdeyleisölle [Valente *et al.*, 2008]. Aste kuvaa myös solmusta lähtevien sellaisten polkujen lukumäärän, joiden pituus on (vähintään) yksi [Borgatti, 2005]. Fortunaton [2010] mukaan verkostojen solmujen astejakauma on laaja ja sisältää potenssilakia noudattavan pitkän hännän. Tämän perusteella verkostossa esiintyy paljon solmuja, joilla on pieni asteluku ja joitakin solmuja, joilla on suuri asteluku. Aste kuvaa lokaalia keskeisyyttä [Scott, 2000], eli sen avulla voidaan arvioida esimerkiksi solmun välitöntä vaikutusvaltaa tai toisaalta riskiä verkostossa [Borgatti, 2005]. Pelkästään solmun yhteyksien määrä ei kuitenkaan kuvaa keskeisyyttä, vaan jotkin yhteydet saattavat olla tärkeämpiä kuin toiset [Hansen *et al.*, 2011]. Lisäksi aste ei mittarina mitenkään huomioi verkoston globaalia rakennetta. Esimerkiksi solmulla saattaa olla paljon yhteyksiä, mutta se ei sijaitse verkostossa sellaisessa positiossa, jossa se saavuttaisi uutta informaatiota nopeasti [Opsahl *et al.*, 2010]. Aste ei keskeisyysmittarina myöskään

huomioi solmujen epäsuoria yhteyksiä mahdollisten vierussolmujen yhteyksien kautta [Landherr *et al.*, 2010]. Freeman [1979] suosittelee asteen käyttöä erityisesti kommunikointiaktiivisuuden tutkimisessa.

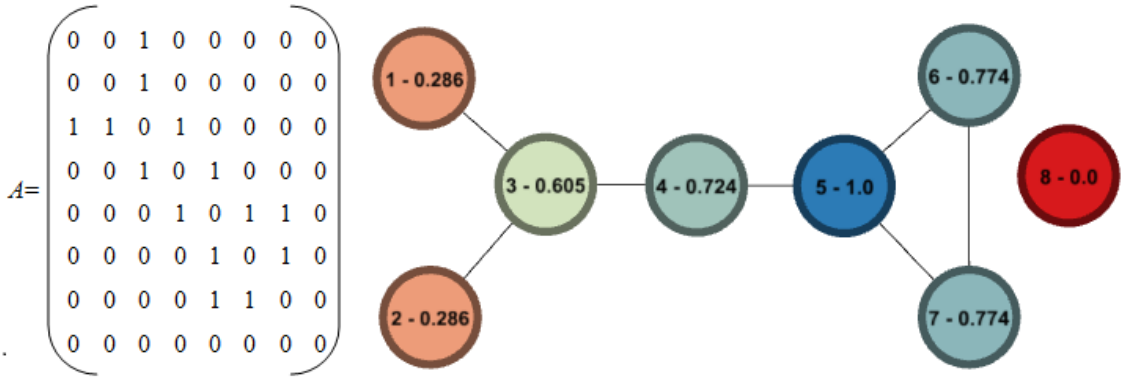
Suunnattujen graafien lähtöasteen avulla voidaan analysoida esimerkiksi solmun aktiivisuutta tai seurallisuutta, ja tuloaste voi kuvata solmun suosiota verkostossa [Opsahl *et al.*, 2010].

Painotettujen graafien solmujen asteet voidaan määritellä laskemalla yhteen solmuun liittyvien särmien painot. Mittaria voidaan kutsua *painotetun asteen* (weighted degree) lisäksi myös *solmun vahvuudeksi* (node strength). Solmun aste 10 voi siten merkitä kymmentä yhteyttä, joiden paino on 1, tai yhtä yhteyttä, jonka paino on 10, tai jotain kombinaatiota näiden väliltä. [Opsahl *et al.*, 2010.] Heikkoutena tässä mittarissa on se, ettei sen avulla voida päätellä solmuun liittyvien särmien ja vierussolmujen lukumäärää.

Ominaisvektoriin (eigenvector) perustuva keskeisyys ratkaistaan matriisilaskennan menetelmillä [Gould, 1967; Bonacich, 1972]. Ominaisvektorikeskeisyytenä käytetään suurinta *ominaisarvoa* (eigenvalue) vastaavan ominaisvektorin arvoja. Ominaisvektori määritellään kaavalla

$$\lambda o = Ao \,,$$

jossa A on graafin vierusmatriisi, λ ominaisarvo ja o ominaisvektori. [Borgatti, 2005.] Kuvassa 14 esitetään esimerkkigraafin vierusmatriisi A ja solmuille lasketut ominaisvektoriarvot. Kuvan arvot on laskettu Gephillä [Bastian *et al.*, 2009], joka skaalaa ominaisvektoriarvot välille 0–1. Suurinta ominaisvektorikeskeisyyttä kuvaa arvo 1, joka on graafissa solmulla 5. Ominaisvektorissa solmun keskeisyys on suhteutettuna sen vierussolmujen keskeisyysarvoihin eli solmun katsotaan olevan ”tärkeä”, jos sen vierussolmut ovat tärkeitä. [Webb and Copsey, 2011.] Kuvasta 14 huomataan, että vaikka solmuilla 3 ja 5 on sama aste, solmu 5 saa suuremman ominaisvektoriarvon, koska sen vierussolmuilla on suurempi aste kuin solmun 3 vierussolmuilla. Borgattin [2005] mukaan ominaisvektori soveltuu hyvin kuvaamaan vaikutusvaltaprosesseja. Solmulla saattaa olla vaikutusvaltaa verkostossa joko suoraan tai epäsuoraan yhteyksiensä kautta [Valente *et al.*, 2008].



Kuva 14. Graafin vierusmatriisi *A* ja solmujen ominaisvektoriarvot.

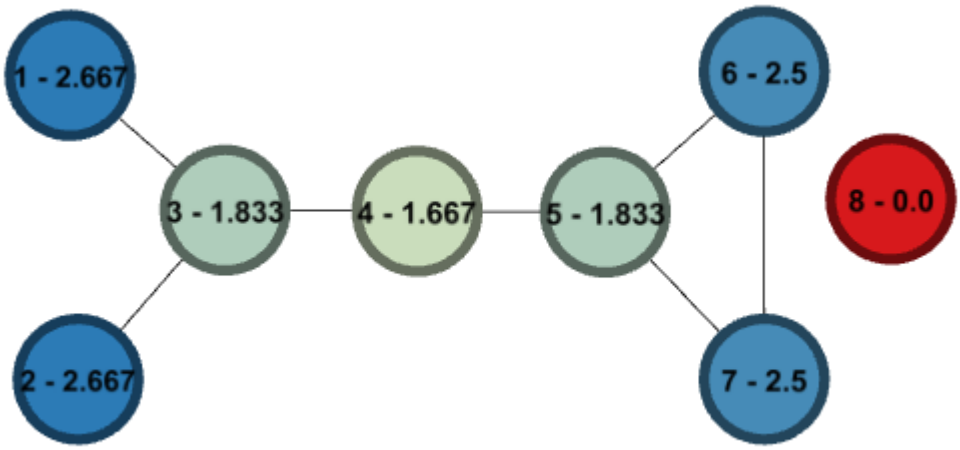
Läheisyys (closeness) on keskiarvo solmun geodeesisten polkujen pituuksista kaikkiin muihin saman komponentin solmuihin. Solmun v_i läheisyys voidaan laskea kaavalla

$$closeness(v_i) = \frac{\sum_{v_j \in V} d_G(v_i, v_j)}{n-1}, \qquad v_i \neq v_j, n \geq 2,$$

jossa $d_G(v_i,v_j)$ on solmujen v_i ja v_j välinen geodeesinen etäisyys ja n yhtenäisen komponentin solmujen lukumäärä [Okamoto *et al.*, 2008]. Esimerkiksi kuvan 15 solmun 4 läheisyys saadaan laskemalla

$$closeness(4) = \frac{2+2+1+1+2+2}{6} = \frac{10}{6} \approx 1{,}667,$$

joka on paras solmun läheisyysarvo esimerkkigraafissa. Läheisyys kuvaa globaalia keskeisyyttä [Scott, 2000], koska solmun v_i etäisyys lasketaan myös niihin komponentin solmuihin, joihin solmulla v_i ei ole suoraa yhteyttä. Matala läheisyysarvo kertoo, että solmu on suoraan yhteydessä tai varsin lähellä muita komponentin solmuja. Siten voidaan päätellä, että kyseiset solmut myös saavuttavat uutta ja arvokasta informaatiota aikaisemmin kuin muut solmut [Borgatti, 2005]. Freemanin [1979] mukaan läheisyydellä voidaankin määritellä solmun itsenäisyys tai tehokkuus, koska suorat yhteydet muihin solmuihin vähentävät informaation vääristymistä ja riippuvuutta muista solmuista. Toisaalta matalat läheisyysarvot voivat kuvata solmuja, jotka ovat ensimmäisinä tartuntatautien saavutettavissa.



Kuva 15. Graafissa solmujen läheisyysarvot.

Sen sijaan korkea läheisyysarvo kuvaa, että komponentin reuna-alueella sijaitsevan solmun tulee muodostaa paljon yhteyksiä luodakseen suhteita etäällä olevien solmujen kanssa [Hansen *et al.*, 2011]. Täten läheisyyttä voi ajatella enemmänkin ”etäisyyden” mittarina. Läheisyysarvo voidaan laskea vain yhtenäisten graafien solmuille, joissa kaikki solmut ovat yhteydessä toisiinsa [Borgatti, 2005]. Epäyhtenäisten graafien irrallisille solmuille, joilla ei ole yhteyksiä muihin solmuihin, voidaan kuitenkin antaa jokin keinotekoinen läheisyysarvo [Valente *et al.*, 2008]. Esimerkiksi kuvassa 15 irralliselle solmulle 8 on annettu läheisyysarvo 0.

Läheisyysarvosta voidaan myös käyttää käänteislukua, jolloin suurempi arvo viittaa keskeisempään solmuun. Tällöin solmun v_i läheisyys voidaan laskea kaavalla

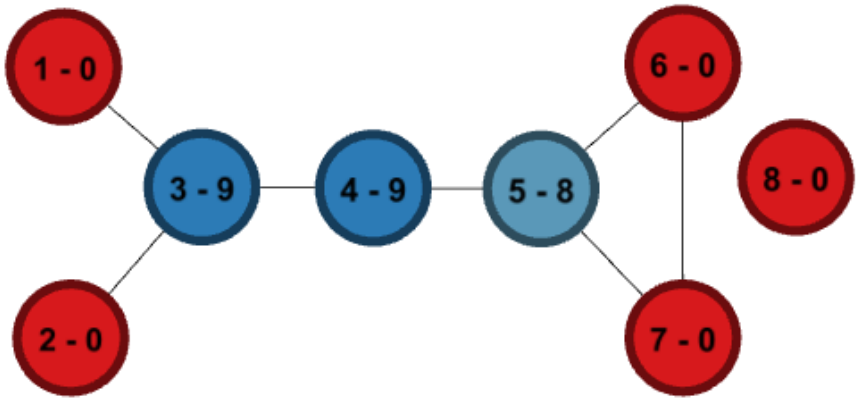
$$closeness(v_i) = \frac{1}{\sum_{v_j \in V} d_G(v_i, v_j)}, \qquad v_i \neq v_j,$$

jossa $d_G(v_i, v_j)$ on solmujen v_i ja v_j välinen geodeesinen etäisyys [Brandes, 2001].

Freemanin [1979] esittelemä mittari *solmun välillisyys* (vertex betweenness) kertoo, kuinka usein tietty solmu voidaan löytää tarkasteltaessa kaikkia solmuparien välisiä geodeesisia polkuja. Solmun v_i välillisyys voidaan laskea kaavalla

$$betweenness(v_i) = \sum_{v_h \in V} \sum_{v_j \in V} \frac{g(v_h, v_i, v_j)}{g(v_h, v_j)}, \qquad v_h \neq v_i \neq v_j,$$

jossa $g(v_h, v_j)$ on geodeesisten polkujen lukumäärä solmusta v_h solmuun v_j ja $g(v_h, v_i, v_j)$ näistä niiden geodeesisten polkujen lukumäärä, jotka kulkevat solmun v_i kautta. Solmun välillisyysmittari ottaa huomioon kaikki mahdolliset lähtö- ja maalisolmujen kombinaatiot. [Borgatti, 2005.] Esimerkiksi kuvan 16 solmu 4 on sellaisten geodeesisten polkujen varrella, jotka kulkevat solmuista 1, 2 ja 3 solmuihin 5, 6 ja 7. Täten solmun 4 välillisyydeksi saadaan 9.



Kuva 16. Graafissa solmujen välillisyysarvot.

Hansen ja muut [2011] mainitsevat välillisyyden yhteydessä myös käsitteen ”*silta-arvo*” (”bridge” score), joka kuvaa sitä, kuinka häiritsevää tietyn solmun poistaminen olisi muiden verkoston yhteyksien kannalta. Välillisyys sopii Freemanin [1979] mukaan erityisesti kommunikoinnin kontrolloinnin tutkimiseen. Kahden toimijan välinen kommunikointi ja vuorovaikutus riippuvat siis niiden välissä olevista toimijoista silloin, kun kyseisten toimijoiden välillä ei ole suoraa yhteyttä [Landherr *et al.*, 2010]. Opsahlin ja muiden [2010] mukaan välillisyyden eräs heikkous liittyy verkostojen yleiseen ominaisuuteen, jonka mukaan suuri osa solmuista ei ole minkään geodeesisen polun varrella, jolloin nämä kaikki solmut saavat välillisyysarvon 0.

Kuten taulukon 1 solmujen keskeisyysarvoista huomataan, eri keskeisyysmittarit järjestävät solmut erilaiseen keskeisyysjärjestykseen. Esimerkiksi solmu 4 on keskeisen positionsa ansiosta kärkisijalla läheisyyskeskeisyydessä ja jaetulla kärkisijalla välillisyyskeskeisyydessä, vaikka se ei yllä keskeisimpien joukkoon

aste- tai ominaisvektorikeskeisyyden perusteella. Ainoastaan irrallinen solmu 8 jää kaikilla mittareilla laskettuna samalle sijalle eli viimeiseksi, koska sillä ei ole yhteyksiä muihin solmuihin.

Taulukko 1. Yhteenveto esimerkkigraafin keskeisyysarvoista.

Aste		**Ominaisvektori**		**Läheisyys**		**Välillisyys**	
arvo	solmut	arvo	solmut	arvo	solmut	arvo	solmut
3	3, 5	1	5	1,667	4	9	3, 4
2	4, 6, 7	0,774	6, 7	1,833	3, 5	8	5
1	1, 2	0,724	4	2,5	6, 7	0	1, 2, 6, 7, 8
0	8	0,605	3	2,667	1, 2		
		0,286	1, 2	0	8		
		0	8				

7.3. Muita solmukohtaisia mittareita

Myös vierussolmujen *astelukujen korrelaatioita* (degree correlations) voidaan analysoida ja vaikkapa selvittää, ovatko suuriasteiset solmut yhteydessä ennemmin samankaltaisten kuin pieniasteisten solmujen kanssa. Lisäksi erityyppisiä linkkien muodostamia malleja voidaan tutkia. [Webb and Copsey, 2011.]

Solmuparia koskeva *särmien yhtenäisyysaste* (edge connectivity) on pienin määrä särmiä, jotka tulee poistaa solmuparin erottamiseksi siten, että solmuparin välillä ei enää ole polkua.

Klusterointikerroin (clustering coefficient) lasketaan jakamalla solmun vierussolmujen välisten yhteyksien lukumäärä maksimaalisella vierussolmujen välisten yhteyksien määrällä [Hansen *et al.*, 2011]. Solmun v_i klusterointikerroin voidaan laskea kaavalla

$$clust\_coeff(v_i) = \frac{E_i}{\frac{1}{2}k_i(k_i - 1)} ,$$

jossa E_i on solmun v_i vierussolmujen välisten särmien ja k_i vierussolmujen lukumäärä [Webb and Copsey, 2011]. Klusterointikerroin on yhteisön tiheysarvon kaltainen mittari, mutta soveltuu paremmin egosentrisiin verkostoihin. Erityisesti sen avulla voidaan mitata 1,5-asteisten verkostojen tiheyttä. Klusterointikerroin on korkea, kun tällaisen verkoston yhteydet ovat tiheässä. Esi-

merkiksi jos kaikki käyttäjän Facebook-ystävät ovat myös ystäviä keskenään, käyttäjän klusterointikerroin on korkea. [Hansen *et al.*, 2011.] Tällaisessa täydellisessä graafissa, jossa kaikki solmut ovat yhteydessä keskenään, jokaisen solmun klusterointikerroin on 1. Käytännössä klusterointikerroin on kuitenkin yleensä 0,1–0,5. Klusterointikerroin kuvaa todennäköisyyttä, että solmun vierussolmut ovat myös toistensa naapureita. [Girvan and Newman, 2002.] Koko graafin klusterointikerroin saadaan laskemalla keskiarvo kaikkien graafin solmujen klusterointikertoimista [Cheng *et al.*, 2008].

Olemassa olevien solmujen lisäksi voidaan tutkia puuttuvia solmuja. Puuttuva silta on rakenteellinen aukko verkostossa [Hansen *et al.*, 2011], eli yhteisöjen väliltä puuttuu niitä yhdistävä solmu.

8. Aikaisempia tutkimuksia keskeisyydestä ja YouTubesta

Sosiaalisten verkostojen solmujen keskeisyydestä on tehty lukuisia tutkimuksia. Useissa tutkimuksissa analysoidaan keskeisyysmittareiden ja verkostojen rakenteellisten ominaisuuksien välisiä suhteita, mutta joissakin tutkimuksissa otetaan huomioon myös solmujen sovellusaluekohtaisia ominaisuuksia. Seuraavaksi kuvataan joitakin keskeisyysmittareihin liittyviä tutkimuksia. Koska tämän tutkielman tutkimusaineistona käytetään YouTube-dataa, perehdytään myös joihinkin YouTubesta raportoituihin tutkimuksiin.

8.1. Tutkimuksia keskeisyysmittareista

Valente ja muut [2008] ovat tutkineet yleisten keskeisyysmittareiden (aste, välillisyys, läheisyys ja ominaisvektori) välisiä riippuvuuksia. He muodostivat mittareista erilaisia symmetrisoituja ja suunnattuja versioita ja testasivat niitä 58 sosiaalisessa verkostossa. Tutkimuksissa havainnoitiin keskeisyyskorrelaatioiden lisäksi verkostojen sosiometrisiä ominaisuuksia (tiheyttä, vastavuoroisuutta, sentralisaatiota ja komponenttien määrää) ja niiden mahdollisia vaikutuksia keskeisyysarvoihin.

Borgatti [2005] on tutkinut keskeisyysmittareita suhteessa *verkoston virtaukseen* (network flow) eli erilaisiin liikennetyyppeihin, joita solmujen välillä vallitsee. Liikennetyyppejä hänen tutkimuksissaan ovat sähköpostin, huhujen, asenteiden, tartuntatautien, rahan, hyödykkeiden ja pakettien kulku. Verkoston virtausta analysoitaessa ei voida käyttää yleisiä keskeisyysmittareita sellaisenaan, koska liikenne ei aina kulje lyhintä polkua pitkin ennalta määriteltyyn maalisolmuun. Lisäksi joissakin solmuissa saatetaan vierailla useita kertoja virtauksen aikana. Esimerkiksi huhu saattaa kulkea saman solmun kautta monesti, mutta tartuntatauti vain kerran taudista muodostuvan immuniteetin vuoksi. Täten solmun tärkeyttä verkostossa ei voida määritellä huomioimatta verkoston virtausta ja sen tyyppiä. [Borgatti, 2005.]

Opsahl ja muut [2010] laajensivat yleisiä keskeisyysmittareita soveltumaan paremmin painotettujen graafien analysointiin ja käyttivät mittareita tutkiessaan 32 tutkijan välistä sähköistä viestien vaihtoa. Heidän mukaansa on oleellista huomioida särmien painojen lisäksi myös niiden määrä painottaen jompaakumpaa tutkimuskohteen mukaan. Geodeesisen etäisyyden laskemisessa huomioitiin polun varrella olevien solmujen määrän lisäksi särmien painot. Lyhin

polku ei nimittäin aina merkitse nopeinta reittiä, vaan esimerkiksi informaatio saattaa kulkea nopeammin vahvoja yhteyksiä pitkin, vaikka polku ei olisikaan solmujen välillä lyhin mahdollinen. Opsahlin ja muiden [2010] mukaan heidän esittelemänsä mittarit painotetuille verkostoille soveltuvat suoraan tietoverkostojen, kuten informaatio- ja neuvontaverkostojen käyttöön. Heikkoutena kyseisissä mittareissa on heidän mukaansa oletus, että särmien painot esitetään suhdeasteikollisina arvoina. Lisäksi voi olla vaikeaa määritellä, halutaanko painottaa enemmän särmien painoja vai määrää, esimerkiksi onko parempi olla useita heikkoja yhteyksiä vai muutama vahva yhteys. [Opsahl *et al.*, 2010.]

8.2. Keskeisyyteen liittyviä tutkimuksia YouTube-datasta

Kang ja muut [2012] painottavat verkostoon liittyvän semantiikan huomioonottamista keskeisyyttä analysoitaessa. Solmujen keskeisyysmittarit kuvaavat nimittäin vain verkostojen rakenteellisia ominaisuuksia, jolloin esimerkiksi tiettyjen ominaisuuksien leviäminen verkostossa jää huomioimatta. Verkoston semantiikalla tarkoitetaan esimerkiksi solmujen ja särmien ominaisuuksia tai painotuksia. Kang ja muut [2012] ehdottavatkin, että solmujen keskeisyys tulisi laskea sekä graafin rakenteellisten ominaisuuksien että esittelemänsä solmujen *diffuusiokeskeisyyden* (diffusion centrality) mukaan. Diffuusiokeskeisyys kuvaa solmujen keskeisyyden graafin rakennetta, semantiikkaa ja *diffuusiomallia* (diffusion model) käyttäen. Diffuusiomalli kuvaa, miten jotkin solmujen ominaisuudet ”monistuvat” verkostossa. Solmu saattaa esimerkiksi omaksua jonkin käyttäytymistavan, jos tarpeeksi moni sen vierussolmuista omaa kyseisen tavan. Käyttäytymistapa voidaan omaksua solmulle luontaisten ominaisuuksien perusteella, eikä verkoston rakenteen mukaan. Diffuusiomalleja voidaan soveltaa esimerkiksi jonkin tuotteen omaksumisen tai taudin leviämistapojen tutkimiseen. Kang ja muut [2012] vertailivat diffuusiokeskeisyyttä ja perinteisempiä keskeisyysmittareita YouTube-datalla. Diffuusio-ominaisuutena he käyttivät YouTuben tiettyihin ryhmiin kuulumista, joten tutkimus on tältä osin vanhentunut, koska ryhmäominaisuus on poistettu palvelusta. Heidän näkemyksensä verkoston semantiikan huomioonottamisesta keskeisyyttä arvioitaessa on kuitenkin yhä merkittävä.

Kulkarni ja Devetsikiotis [2010] käyttivät YouTuben videodataa tutkimuksissaan, joissa analysoitiin videoiden välisiä suhteita eri ajanjaksoina ja arvioitiin videoiden tärkeyttä keskeisyysmittareiden (aste, läheisyys ja välillisyys) perusteella. Tutkimusten tarkoituksena oli osoittaa, että tiettyä keskeisyysmenetelmää käyttämällä videoiden välimuistikuormaa voidaan vähentää merkittävästi. Kulkarnin ja Devetsikiotisin [2010] mielestä videoiden välimuistin tulisi perus-

tua sekä yksittäisen videon että siihen liittyvien videoiden suosioon. Videoiden väliset suhteet ovat siten pääosassa heidän tutkimuksiaan. Kulkarni ja Devetsikiotis [2010] toteavat, että läheisyys oli keskeisyysmittareista paras videoiden valinnassa. Välillisyys sen sijaan osoittautui heikoksi mittariksi merkittävimpien videoiden valinnassa, mutta saattaisi heidän mukaansa toimia paremmin, jos koko graafidata olisi saatavilla.

8.3. Tutkimuksia YouTuben verkostoista ja rakenteesta

Myös Cha ja muut [2007] ovat tutkineet videoiden välimuistimenetelmiä. Erityisesti he keskittyivät tutkimuksissaan kuitenkin videoiden suosion elinkaareen ja videoiden katselun määrään suhteessa videoiden ikään. Tutkimuksissa huomattiin, että 80 % tiettynä päivänä katsotuista YouTuben videoista oli yli kuukauden vanhoja. Videoiden katseluun liittyen voidaan muodostaa ja tutkia käyttäjien *katselumalleja* (view patterns). Cha ja muut [2007] havaitsivat, että 10 % suosituimmista videoista saavuttaa lähes 80 % kaikista katselumääristä. Tätä havaintoa voisi heidän mukaansa hyödyntää välimuistin tallennusmenetelmissä.

Cheng ja muut [2008, 2013] ovat tutkineet YouTuben videoverkostoja ja videoihin liittyvää statistiikkaa. He huomasivat, että YouTuben videoverkostojen rakenteella on selkeä pienen maailman ominaispiirre. Tämän mukaan videoiden välillä on voimakkaita riippuvuuksia. Pienen maailman ominaisuus todettiin vertaamalla videoverkostojen ja vastaavien satunnaisten graafien klusterointikertoimia ja solmujen välistä keskimääräistä etäisyyttä. Graafi vastaa pienen maailman ilmiötä, jos klusterointikerroin on korkea, mutta solmujen välinen keskimääräinen etäisyys pieni. Chengin ja muiden [2008] mukaan solmujen välinen pieni etäisyys on odotettavissa, koska YouTuben toisiinsa liittyvät videot ovat löydettävissä käyttäjien luomien tunnisteiden sekä videoiden nimi- ja kuvailutietojen kautta.

Myös YouTube-käyttäjien välisiä yhteyksiä ja yhteydenpitoa on tutkittu. Spathis ja Gorcitz [2011] tutkivat yhteisöllisyyteen pohjautuvien ominaisuuksien suhdetta videoiden suosioon. Heidän mukaansa tällainen tutkimus voi parantaa käyttäjille tarjottavia toimintoja, kuten hakuominaisuuksia, personoituja suositteluja ja informaation jakamista. Tutkimus on osittain vanhentunut, koska siinä keskityttiin YouTuben ryhmiin liittyvään yhteisöllisyyteen, ja tämä toiminto ei enää ole palvelussa mahdollista. Tutkimuksessa kuitenkin huomattiin, että YouTuben käyttäjät ovat sitoutuneempia silloin, kun he keskittyvät vain muutamiin tiettyihin aiheisiin. Lisäksi havaittiin, että kun käyttäjät lataavat vi-

deoitaan spesifisen kategorian alle, he saavat paljon tunnustusta ja huomiota osakseen. Spathis ja Gorcitz [2011] vertaavat käyttäjän saamaa tunnustusta perinteisten yritysten tekemiin brändäyksiin.

YouTubea ei aina mielletä yhteisönä. Rotman ja muut [2009] ovat tutkineet YouTube-käyttäjien yhteisöllisyyden tuntemuksia ja niiden yhteyttä mahdollisiin eksplisiittisiin siteisiin käyttäjien välillä. He analysoivat videoita ja niiden kommentteja sekä käyttäjien välisiä yhteyksiä. Eksplisiittiset suhteet osoittautuivat melko harvinaisiksi käyttäjien kesken, mutta käyttäjillä oli kuitenkin tuntemuksia yhteisöön kuulumisesta YouTubessa. Tämä osoittaa, että kaikki yhteisöllisyyden piirteet eivät välttämättä näy suurten järjestelmien verkostorakenteessa, joten tarvitaan sekä kvantitatiivista että kvalitatiivista dataa ja analysointia verkkoyhteisöjen luonteen ymmärtämiseksi.

YouTuben käyttäjädataa käytettiin osana Misloven ja muiden [2007] tutkimuksia. He tutkivat internetin sosiaalisten verkostojen rakenteellisia ominaisuuksia. Graafin rakenteiden perusteella voidaan suunnitella uusia verkkoyhteisöpalveluja, evaluoida olemassa olevia sovelluksia ja ymmärtää niiden vaikutuksia internetissä. Mislove ja muut [2007] havaitsivat tutkimuksissaan, että suunnatuissa sosiaalisissa verkostoissa vallitsee suuri vastavuoroisuus. Solmuilla, joilla on hyvin korkea lähtöaste, todettiin olevan myös hyvin suuri tuloaste. Tutkimuksissa havaittiin myös, että vaikka verkostot kasvavat nopeasti, niiden perusrakenne ei muutu merkittävästi. Verkostoissa suurimmalla osalla solmuja on pieni aste ja vain muutamalla solmulla on merkittävän korkea aste. Nämä korkeaasteiset solmut kuitenkin muodostavat verkostoissa tärkeän ydinosan, jota tarvitaan pitämään yllä yhteyksiä muihin verkoston solmuihin. Kyseisillä solmuilla voi siten olla merkittävää vaikutusvaltaa informaation kulkuun, luotettavuuteen ja verkoston haavoittuvuuteen liittyen.

9. Tutkimus keskeisistä solmuista YouTube-videoverkostoissa

Kuten luvun 8 perusteella huomataan, YouTube-dataan liittyviä keskeisyystutkimuksia on raportoitu verrattain vähän. Rotmanin ja Golbeckin [2011] mukaan sosiaalisten verkostojen tutkimuksissa on keskitytty enimmäkseen YouTubea selkeämpiin sosiaalisiin verkostoihin, kuten Facebookiin ja Twitteriin. Jotkin YouTubeen liittyvät tutkimukset ovat myös osittain vanhentuneita, koska YouTube on muuttanut joitakin yhteisöllisyyteen liittyviä ominaisuuksia palvelussaan. Esimerkiksi YouTuben sisäiset käyttäjistä koostuneet ryhmät ovat poistuneet, ja lisäksi käyttäjät eivät enää voi lisätä toisiaan ystävikseen. Myös mahdollisuus vastata videoon toisella videolla on poistettu ominaisuuksista.

Tässä tutkimuksessa selvitetään, minkälaisia graafeja YouTubesta haetun videodatan perusteella muodostuu, onko graafeissa rakenteellisesti samankaltaisia elementtejä ja löytyykö graafeista yhteisöjä. Haettu data on rajoitettu kuuteen eri aiheeseen, joista kerrotaan tarkemmin kohdassa 9.2. YouTube-videoiden suosiota arvioidaan usein niiden ominaisuuksien, kuten katselukertojen ja annettujen arvostelujen perusteella. Tutkimuskysymyksinä on selvittää, mitkä videot ovat keskeisiä keskeisyysmittareiden perusteella, ja vastaavatko nämä videot katselukertojen perusteella suosittuja videoita. Lisäksi tässä tutkimuksessa esitellään keskeisyysarvojen ja katselukertojen perusteella muodostettu pisteytysmenetelmä, jonka perusteella esitellään videoverkostojen keskeisimmät solmut. Tässä tutkimuksessa tutkitaan myös, miten solmukohtaiset keskeisyysmittarit (aste, painotettu aste, ominaisvektori, läheisyys ja välillisyys) korreloivat sovellusalueen attribuuttidatan (katselukerrat ja arvosana) kanssa. Myös eri keskeisyysarvojen välillä olevia ja attribuuttien välisiä korrelaatioita tarkastellaan.

9.1. Yleistä YouTube-datasta

Tässä tutkimuksessa käytettiin YouTubesta haettua dataa. YouTube on vuonna 2005 perustettu sosiaalinen videopalvelusivusto, josta on muodostunut maailman suosituin videoiden jakamiseen erikoistunut palvelu [Rotman and Golbeck, 2011]. Sivuston videot perustuvat *käyttäjien tuottamaan sisältöön* (user-generated content). YouTuben tilastotietojen mukaan sivustolla vierailee yli miljardi käyttäjää joka kuukausi, ja videopalveluun ladataan 100 tuntia videosisältöä joka minuutti. Tämän lisäksi käyttäjät tekevät miljoonia *tilauksia* (subscription) päivittäin. [YouTube, 2014.] Tilauksella tarkoitetaan suunnattua, epä-

symmetristä yhteystyyppiä, jossa tilaaja eli videoiden katsoja saa ilmoitukset tilaamiensa käyttäjien uusista videoista omalle *kanavalleen* (channel). Kanavalla tarkoitetaan käyttäjän henkilökohtaista profiilia, jota pystyy jonkin verran muokkaamaan. Esimerkiksi omat tilaukset voi asettaa yksityisiksi, jolloin muut käyttäjät eivät pääse niitä näkemään. YouTuben suosio perustuu Chengin ja muiden [2013] mukaan sekä sisällöiltään monipuolisiin videoihin että sosiaalisten verkostojen muodostumiseen. Sosiaalisilla verkostoilla saattaa heidän mukaansa olla jopa suurempi vaikutus sivuston suosioon.

YouTuben verkostoja voidaan tarkastella sekä videoiden että käyttäjien näkökulmasta. Videoverkostot voivat olla samankaltaisiin sisällön kuvailutietoihin, kuten tunnisteisiin, nimikkeisiin, kuvauksiin ja kategorioihin perustuvia. Valmiiksi määriteltyjä kategorioita on lukuisia, esimerkiksi musiikki, komedia, elokuvat ja viihde, kauneus ja muoti, urheilu, tiede ja koulutus, uutiset ja politiikka sekä tee-se-itse-ohjeet. Cheng ja muut [2013] havaitsivat tutkimuksissaan, että puolet YouTuben videoista liittyvät joko musiikki- tai viihdekategoriaan. Videon lataaja voi liittää videoon avainsanoja tunnisteiksi ja liittää siihen linkkejä muihin videoihin. Videot eivät siten ole toisistaan riippumattomia, koska niitä voidaan selata ja katsella linkkejä seuraamalla [Cheng *et al.*, 2013]. Videoita voidaan jakaa myös YouTube-palvelun ulkopuolella, esimerkiksi lisäämällä videolinkkejä sähköpostiviesteihin tai muihin sosiaalisen median sivustoihin.

Videoverkostojen peruselementteihin kuuluvat myös videoihin liitetyt kommentit, joita käyttäjät voivat lisätä videokohtaisesti. Tekstikommentit ovat Rotmanin ja muiden [2009] mukaan YouTuben käytetyin kommunikointikeino. Kommentit vaihtelevat Rotmanin ja Golbeckin [2011] mukaan lukukelvottomista merkinnöistä syvällisiin keskusteluihin. Suurimpaan osaan kommenteista ei kuitenkaan vastata. Kommentoinnilla voidaan pyrkiä keskustelemaan videon sisällöstä ja osoittamaan hyväksyntää tai vastenmielisyyttä joko videota tai sen tekijää kohtaan. YouTube-videoiden lataajat pitävät saamiaan kommentteja keinona saavuttaa merkittävyyttä ja näkyvyyttä muiden käyttäjien keskuudessa, ja omien videoiden kommentteja sekä kommentoijia arvostetaan paljon. [Rotman *et al.*, 2009.] Rotman ja Golbeck [2011] toteavat, että YouTubessa on erilaisia suosion ja aktiivisuuden tasoja. Tällä he tarkoittavat sitä, että jotkin videot ovat suosittuja, mutta eivät herätä keskustelua ja toisin päin. Susarlan ja muiden [2012] mukaan YouTube-videon katselu vastaa kuluttajan tekemää valintaa jonkin uuden tuotteen suhteen – katsoja kohtaa valinnan, katsoako video vai ei.

Kommentteja voidaan tarkastella myös käyttäjäverkoston näkökulmasta, koska kommentit yksilöidään käyttäjätunnuksen perusteella. Käyttäjäverkostosta voidaan tutkia myös tilauksia, jos niitä ei ole asetettu yksityisiksi. Käyttäjäverkostot ovat egosentrisiä ja voivat sisältää sekä eksplisiittisiä että implisiittisiä yhteystyyppejä. Kanavan tilaaminen on esimerkki eksplisiittisestä yhteystyypistä. Implisiittisiä suhteita syntyy esimerkiksi silloin, kun käyttäjät kommentoivat samaa videota. [Rotman and Golbeck, 2011.] Käyttäjien kommentointi johtaa epäsuoraan myös videoverkostojen muodostumiseen, joita tässä tutkimuksessa käytetään esimerkkidatana.

9.2. Käytetyt työkalut ja datan hakeminen

YouTube-data haettiin NodeXL-työkalulla, joka on Microsoft Exceliin (2007, 2010 ja 2013) ladattavissa oleva avoimen lähdekoodin lisäosa [Smith *et al.*, 2010]. NodeXL sisältää työkalun, jonka avulla dataa voidaan hakea suoraan sosiaalisen median sivustoilta, esimerkiksi Twitteristä tai, kuten tässä tutkimuksessa, YouTubesta. NodeXL käyttää videodatan hakemisessa YouTuben tarjoamaa ohjelmointirajapintaa (YouTube Data API). Työkalun avulla voidaan hakusanojen perusteella hakea sekä videoihin että käyttäjiin liittyvää dataa. YouTuben käyttämä Googlen hakukoneen algoritmi täsmäyttää hakusanat kaikkiin YouTube-videoista löytyviin kuvailutietojen merkkijonoihin, eli videoiden nimiin, kuvauksiin, käyttäjätunnuksiin, tunnisteisiin ja kategoriatietoihin. YouTuben ohjelmointirajapinta rajoittaa palauttamiensa videoiden määrää maksimissaan 999:ään [YouTube API, 2014]. Tässä tutkimuksessa videoiden määrä vaihteli aiheesta riippuen ja oli suurimmillaan 500.

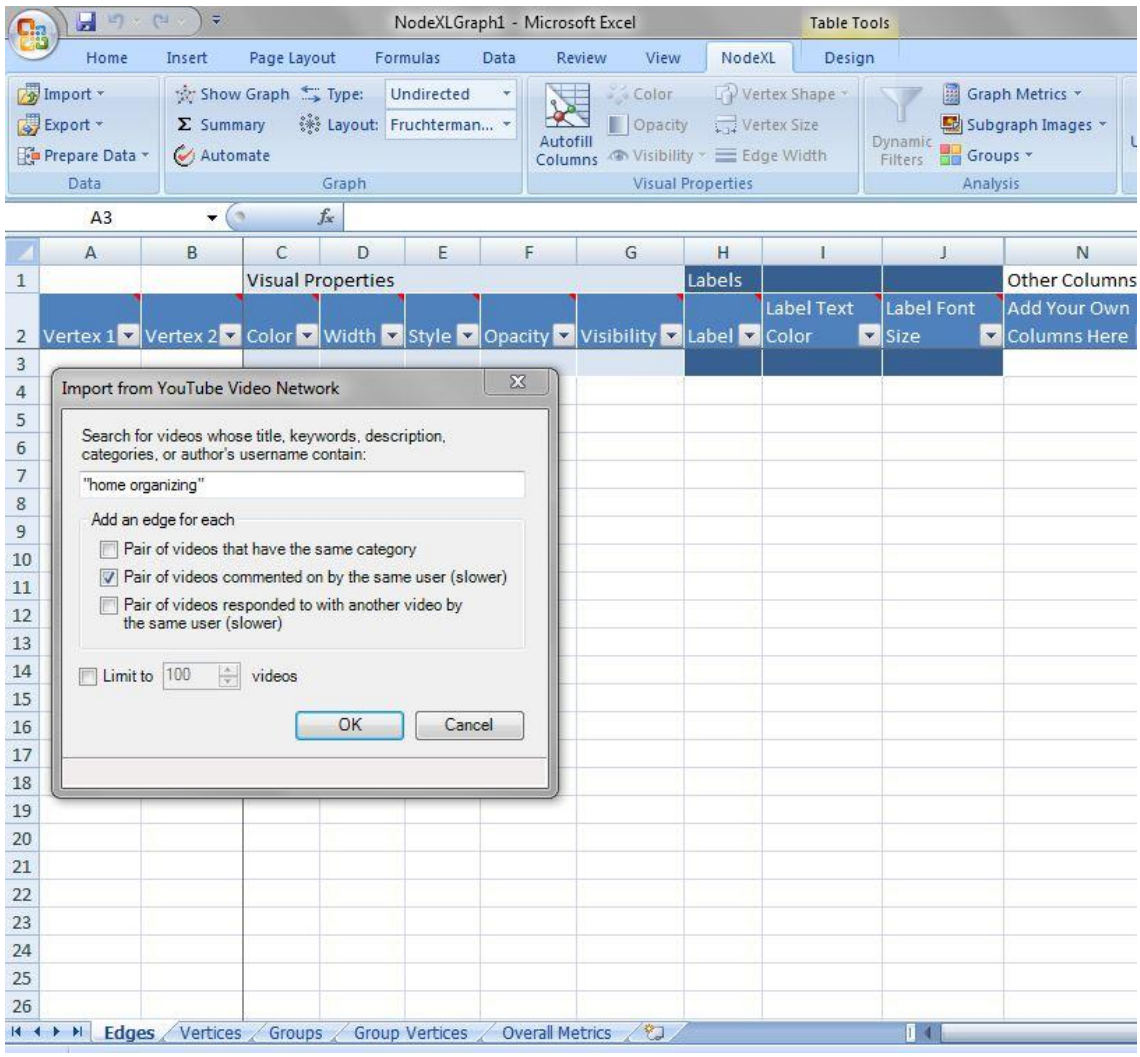
YouTuben videoverkostosta haettiin yhteensä 2271 videota kuudesta erilaisesta aiheesta:

1. Paulo Coelho The Witch of Portobello (255 videota)
2. ”kettlebell workout” (490 videota)
3. “home organizing” (465 videota)
4. 50’s pin up rockabilly makeup tutorial (463 videota)
5. How to cut your own hair (500 videota)
6. ”New York travel tips” (98 videota)

Aiheet valittiin melko erilaisista teemoista ja eri spesifisyysasteista. Esimerkiksi ”home organizing” on hyvin yleinen aihe, joka voi sisältää monenlaisia videoita. Tästä aiheesta löytyykin satojatuhansia videoita YouTube-haulla: n. 488 000 tulosta (31.3.2014). Sen sijaan Paulo Coelhon kirjaan liittyvä ensimmäinen haku on hyvin spesifi, ja YouTubesta löytyy haulla vain joitakin satoja videoita: n. 376 tulosta (31.3.2014). Laajempien aiheiden hakutuloksia pyrittiin rajoittamaan

ja tarkentamaan lainausmerkeillä erotelluilla fraasihauilla, esimerkiksi ”kettlebell workout”.

Videoiden yhteystyypiksi valittiin yhteinen kommentoija, eli kahden videoita kuvaavan solmun välillä on särmä, jos niillä on yhteinen kommentoija. Tämän avulla pyrittiin selvittämään, löytyykö muodostuneista graafeista mahdollisia yhteisen aihepiirin ja kommentoijien muodostamia yhteisöjä. Kuvassa 17 esitetään eräs videodatan hakutilanne NodeXL-työkalulla.



Kuva 17. Videodatan haku NodeXL-työkalulla.

Haettu data sisältää videoon liittyvää metadataa, joka kuvaa videon ominaisuuksia. Osa videoiden ominaisuuksista, kuten katselukertojen määrä ja arvosana, ovat dynaamisia ja saattavat muuttua hyvinkin nopeasti. Tässä tutkimuksessa ominaisuuksia analysoitiin datan keräämishetken mukaisten lukumäärien

mukaan. Videoiden arvosanat ja katselukertojen määrä voivat kuvata sekä videoiden suosiota että niiden saavutettavuutta. Cheng ja muut [2008, 2013] havaitsivat tutkimuksissaan, että videoilla on erilaisia kasvutrendejä, eli videoiden suosio kasvaa eri vauhtia. He löysivät esimerkiksi hyvin paljon katsottuja uusia videoita ja vain vähän katsottuja vanhempia videoita. Vain kirjautunut käyttäjä voi kommentoida tai arvostella videota, joten sekä kommenttien että arvostelujen määrä on huomattavasti alhaisempi kuin katselukertojen määrä. Joillakin videoilla ei ole yhtään kommenttia tai arvostelua. Tämän tutkimuksen videot ovat yhteydessä toisiinsa yhteisten kommentoijien ja yhteisen aiheen kautta. Cheng ja muut [2008, 2013] toteavat, että jos ryhmä videoita on tiiviisti yhteydessä toisiinsa, käyttäjä todennäköisesti katsoo toisen videon tämän ryhmän sisältä ensimmäiseksi katsomansa videon jälkeen.

Jokainen video on yksilöity 11-merkkisellä ID-tunnuksella. Muita attribuutteja ovat videon nimi, lataajan käyttäjätunnus, latausaika, arvosana, katselukerrat, kommenttien lukumäärä ja videon URL-osoite. Taulukossa 2 esitetään yhden tutkimuksessa käytetyn videon attribuutit ja metadata (haettu 31.3.2014).

Taulukko 2. Esimerkki videodatasta.

ID-tunnus	e2qxG-3V94U
Videon nimi	The Experimental Witch
Lataaja	WitchOfPortobello
Latausaika	9.7.2007
Arvosana	4,6666665
Katselukerrat	28 250
Kommenttien lkm	10
URL	http://www.youtube.com/watch?v=e2qxG-3V94U&feature=youtube_gdata_player

YouTube käytti aiemmin videoiden arvioinnissa asteikkoa 1–5. Arviointi muutettiin kuitenkin ”tykkää” ja ”ei tykkää”-muotoiseksi vuonna 2010. Positiivinen arvio vastaa nykyään arvosanan laskemisperiaatteessa numeroa 5 ja negatiivinen arvio numeroa 1. Arvosana saadaan laskemalla yhteen annettuja arvioita vastaavat arvot ja jakamalla tämä summa annettujen arvioiden lukumäärällä.

Data tallennettiin XML-rakennetta noudattavassa GraphML-tiedostoformaatissa ja visualisoitiin Gephillä, joka on graafien visualisointiin ja analysointiin tarkoitettu avoimen lähdekoodin ohjelmistopaketti [Bastian *et al.*,

2009]. Lisäksi Gephillä laskettiin solmujen keskeisyysarvot (aste, painotettu aste, ominaisvektori, läheisyys ja välillisyys) ja muita graafiin liittyviä mittareita. Verkostojen visualisoinnissa käytettiin Gephin Force Atlas -algoritmia, joka tukee solmujen asettelua siten, että ne eivät sijoitu graafiin päällekkäin. Force Atlas -asettelu soveltuu erityisesti kohtuullisen pienten sosiaalisten verkostojen esittämiseen ja tutkimiseen [Gephi Tutorial Layouts, 2014].

9.3. Graafien visualisointi ja analysointi

Videoverkostoista muodostuneet graafit ovat suuntaamattomia. Solmut kuvaavat videoita ja videoparien väliset painotetut särmät kuvaavat, että videoparilla on vähintään yksi yhteinen kommentoija. Särmiä on painotettu yhteisten kommentoijien lukumäärällä. Jokaisesta videoverkostosta esitellään aluksi visuaalinen yleiskuva, josta saadaan alustavaa tietoa graafin rakenteesta. Yleiskuvassa ovat esillä kaikki videoverkoston solmut ja niiden väliset särmät. Särmän paksuuden avulla esitetään särmän painoa. Solmujen koko ja värisävy on skaalattu asteen mukaan siten, että suurimmat ja vaaleimmat solmut ovat suurimpia myös asteeltaan.

Hakusanoilla ”New York travel tips” löytynyt videoverkosto sisälsi 98 solmua, mutta vain yhden särmän, joten tämä data jätettiin keskeisyysmittareiden laskemisen ja tarkemman tarkastelun ulkopuolelle.

Muiden videoverkostojen solmuille lasketut keskeisyysarvot järjestettiin paremmuusjärjestykseen eli läheisyysarvoja lukuunottamatta laskevaan järjestykseen. Läheisyys on laskettu siten, että pienin arvo kuvaa suurinta läheisyyttä. Laskettujen keskeisyysarvojen lisäksi tarkasteluun otettiin mukaan videoiden katselukerrat. Jatkossa keskeisyysarvoilla ja -mittareilla viitataan sekä rakenteellisiin ominaisuuksiin (aste, painotettu aste, ominaisvektori, läheisyys, välillisyys) että katselukertojen määrään. Jokaisesta videoverkostosta esitellään suurimmat keskeisyysarvot noin kymmenen ensimmäisen solmun osalta. Jos usealla solmulla esiintyi sama keskeisyysarvo kuin 10. solmulla, otettiin kaikki samanarvoiset solmut mukaan tarkasteluun. Desimaaliluvut on pyöristetty kolmen desimaalin tarkkuudella.

Arvosanaa ei otettu mukaan keskeisyysarvioihin, koska videon lataaja voi kieltää videonsa arvioinnin. Täten puuttuvaa dataa on melko paljon arvosana-attribuutin kohdalla. Lisäksi huomattiin, että satoja videoita verkostosta oli saattanut saada arvosanakseen täydet 5 pistettä, joten noin kymmentä keskeisintä videota ei tämän mittarin perusteella voida määritellä.

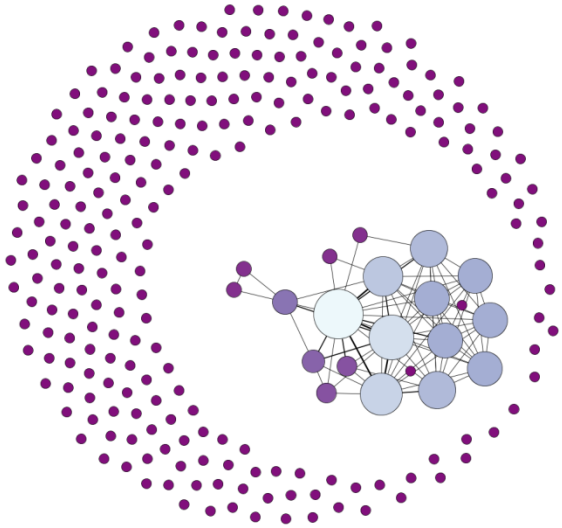
Solmut pisteytettiin keskeisyysarvojen perusteella siten, että suurimman keskeisyysarvon omaava solmu sai kymmenen pistettä. Jos usealla solmulla oli sama keskeisyysarvo, saivat solmut saman pistemäärän kyseisen keskeisyysmittarin suhteen. Läheisyysmittarin osalta tarkastelusta poistettiin ne solmut, jotka kuuluivat hyvin pieniin (1–5 solmua) komponentteihin, koska tarkastelussa haluttiin painottaa koko verkoston kannalta keskeisimpien solmujen löytämistä. Pisteet laskettiin kaikkien valittujen kuuden keskeisyysmittarin perusteella mukaan lukien videoiden katselukerrat. Solmun saamat keskeisyyspisteet laskettiin yhteen. Solmu saattoi siten saada maksimissaan 60 pistettä. Lähempään tarkasteluun valittiin noin kymmenen solmua jokaisesta videoverkostosta suurimpien kokonaispisteiden perusteella.

Korkeimmat kokonaiskeskeisyyspisteet saaneiden solmujen lähempää visuaalista tarkastelua varten koko verkostoa kuvaavasta graafista suodatettiin pois irralliset solmut ja muutamien solmujen muodostamat pienet komponentit. Jokaisesta verkostosta pyrittiin tuottamaan mahdollisimman luettava visuaalinen esitys, johon keskeisimmät solmut on merkitty.

Seuraavaksi tarkastellaan viiden erilaisen aiheen mukaisia videoverkostoja yksi kerrallaan.

1. Videoverkosto 1 (Paulo Coelho The Witch of Portobello)

Videoverkostoa 1 kuvaavassa graafissa on 255 solmua ja 77 painotettua särmää. Graafi muodostuu 237 komponentista, joista suurin sisältää 19 solmua. Loput 236 ovat yksisolmuisia, joten irrallisten solmujen osuus on tässä graafissa hyvin suuri (92,5 %). Yleiskuva videoverkostosta 1 esitetään kuvassa 18. Koko graafin solmujen asteet ovat välillä 0–16. Särmien paino vaihtelee graafissa yhdestä kolmeen. Koko graafin modulaarisuus on 0,133.



Kuva 18. Yleiskuva koko graafista 1 (Paulo Coelho The Witch of Portobello).

Taulukossa 3 esitetään parhaat keskeisyysarvot 8–11 ensimmäisen solmun osalta. Tässä verkostossa välillisyysarvon perusteella otettiin tarkasteluun vain kahdeksan solmua, koska lopuilla graafin solmuilla välillisyysarvo on 0.

Taulukko 3. Keskeisyysarvoja (8–11 ”parasta” solmua graafissa 1).

Aste		**Painot. aste**		**Ominaisvektori**		**Läheisyys**		**Välillisyys**		**Katselukerrat**	
arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu
16	8	26	8	1,000	8	1,111	8	38,417	8	61465	8
14	7	21	7	0,968	7	1,222	7	32,250	3	52610	7
13	6	18	6	0,939	6	1,333	9	17,917	7	28250	6
12	9	15	9	0,894	9	1,389	6	13,750	9	23347	9
11	11	12	11	0,881	11	1,500	11	7,667	6	21804	3
11	17	12	17	0,865	17	1,500	17	4,500	17	15291	18
10	12	11	14	0,849	12	1,556	12	1,750	4	15227	10
10	13	10	12	0,849	13	1,556	13	1,750	11	15084	21
10	14	10	13	0,849	14	1,556	14			14424	65
10	15	10	15	0,849	15	1,556	15			12452	86
10	16	10	16	0,849	16	1,556	16				

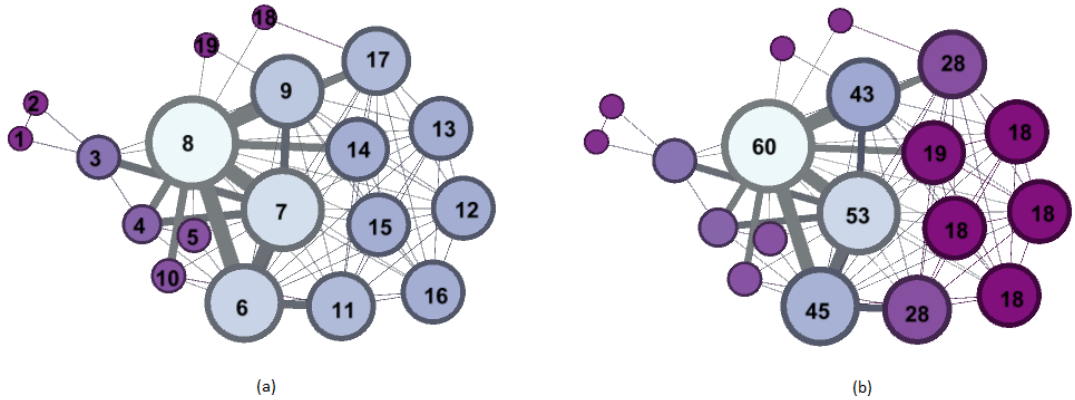
Taulukosta 4 huomataan, että solmu 8 on kaikilla keskeisyysarvoilla arvioituna ensimmäisenä. Täten se myös sai maksimipistemäärän 60. Koska tässä verkostossa vain 19 solmulla on yhteyksiä muiden solmujen kanssa,

pistetaulukon ulkopuolelle jää vain kahdeksan solmua. Taulukossa on tyhjiä soluja, koska kaikki välillisyyden ja katselukertojen perusteella keskeisimpien joukkoon päässeet solmut eivät ole muilla mittareilla laskettuna keskeisimpiä.

Taulukko 4. Solmujen pisteytys (11 keskeisintä graafissa 1).

Solmu	Aste	Painotettu aste	Ominais-vektori	Läheisyys	Välillisyys	Katselu-kerrat	Pisteet yhteensä
8	10	10	10	10	10	10	60
7	9	9	9	9	8	9	53
6	8	8	8	7	6	8	45
9	7	7	7	8	7	7	43
11	6	6	6	6	4		28
17	6	6	5	6	5		28
14	5	5	4	5			19
12	5	4	4	5			18
13	5	4	4	5			18
15	5	4	4	5			18
16	5	4	4	5			18

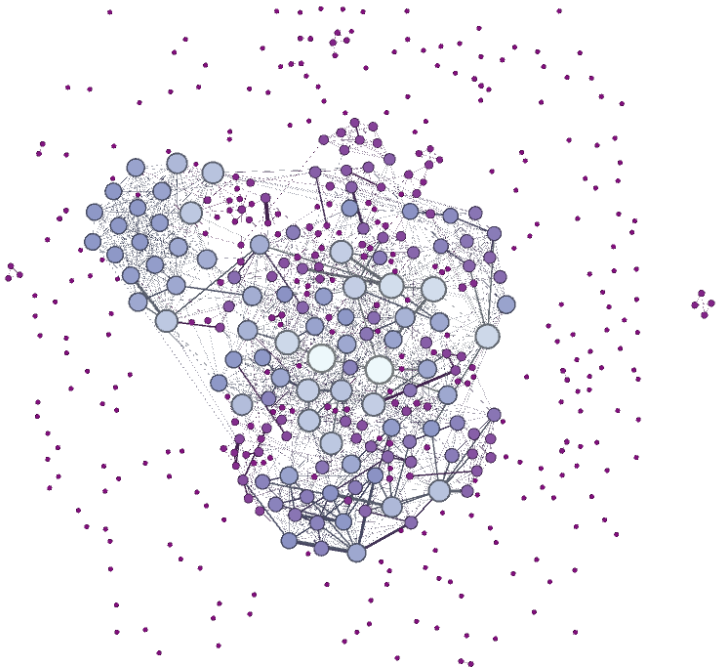
Kuvan 19 graafista on suodatettu pois irralliset solmut, joilla ei ole yhteyksiä muihin solmuihin. A-kohtaan on merkitty kaikki jäljelle jäävän yhtenäisen komponentin solmut. Pisteytyksen mukaan keskeisimmät solmut on kuvattu b-kohdassa, ja kokonaispistemäärät on merkitty solmuihin. Tässä videoverkostossa kaikki keskeisimmät solmut muodostavat täydellisen aligraafin, eli ne ovat kaikki yhteydessä toistensa kanssa. Kaikilla 11 videolla on siten vähintään yksi yhteinen kommentoija.



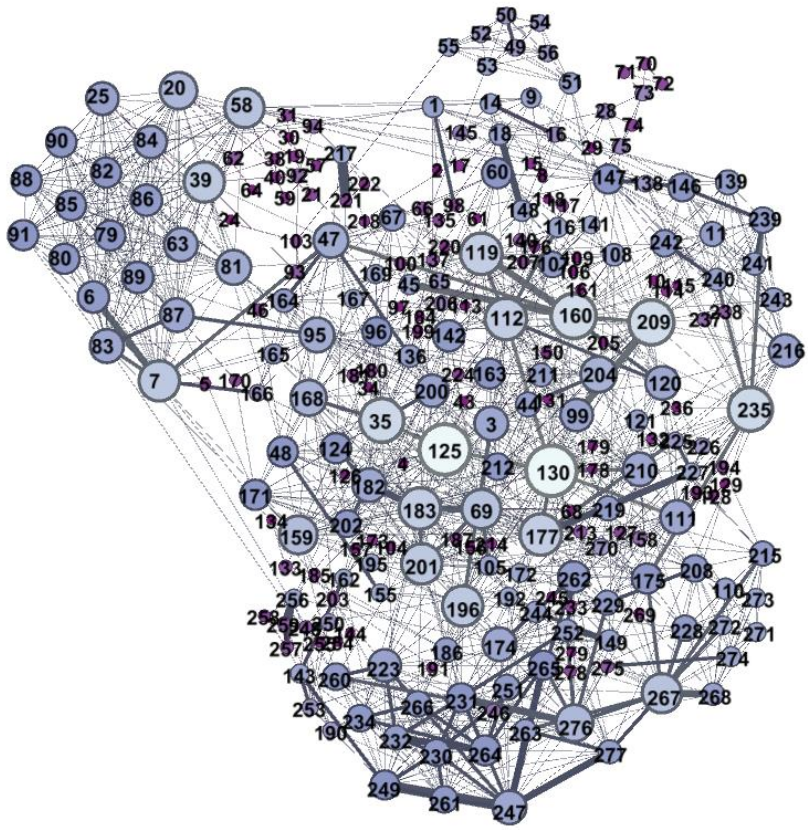
Kuva 19. Videoverkoston 1 (a) suurimman komponentin solmut,
(b) suurimmat kokonaiskeskeisyyspisteet.

2. Videoverkosto 2 (”kettlebell workout”)

Toinen videoverkosto on esitetty kuvassa 20. Graafissa on 490 solmua ja 1349 särmää muodostaen 223 komponenttia, joista suurimmassa on 252 solmua. Suurimmassa komponentissa on siten yli puolet graafin solmuista (n. 51,4 %). Tämä komponentti esitellään kuvassa 21. Irrallisia solmuja on 211 eli 43,1 % koko graafista. Solmujen aste vaihtelee välillä 0–37. Särmien paino on graafissa välillä 1–4. Koko graafin modulaarisuus on 0,585.



Kuva 20. Yleiskuva koko graafista 2 (”kettlebell workout”).



Kuva 21. Suurin komponentti graafissa 2.

Taulukossa 5 esitetään parhaat keskeisyysarvot 10 ensimmäisen solmun osalta, joiden mukaan laskettujen keskeisyyspisteiden perusteella järjestetyt solmut esitetään 10 suurimman osalta taulukossa 6.

Taulukko 5. Keskeisyysarvoja (10 ”parasta” solmua graafissa 2).

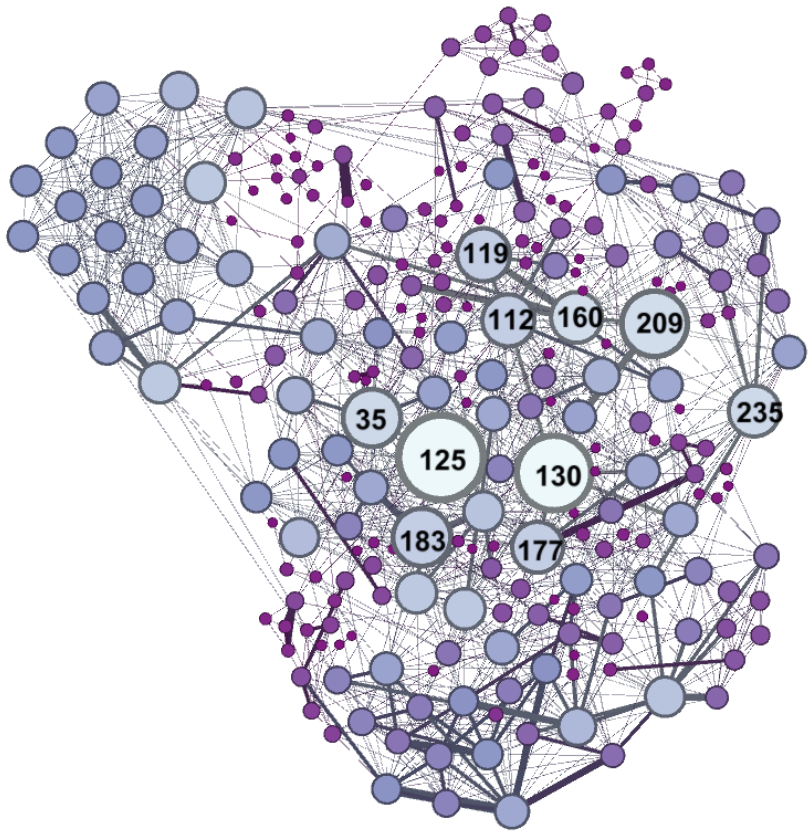
Aste		Painot. aste		Ominaisvektori		Läheisyys		Välillisyys		Katsojamäärä	
arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu
37	125	42	130	1,000	130	2,287	125	1941,688	204	1277265	282
37	130	39	125	0,987	125	2,339	209	1760,838	125	978152	262
32	160	38	160	0,797	35	2,355	130	1627,247	112	804876	171
32	209	36	267	0,788	201	2,359	119	1538,007	39	730233	277
31	35	35	209	0,786	209	2,363	177	1503,368	159	634878	281
31	235	34	35	0,777	183	2,375	69	1429,685	47	590236	201
29	112	34	119	0,754	177	2,378	183	1364,075	235	508851	136
29	119	34	183	0,740	7	2,402	112	1348,592	58	496903	1
29	177	34	235	0,738	159	2,406	35	1330,826	7	485870	183
29	183	34	247	0,727	196	2,410	47	1282,737	60	473596	260

Suurimman katsojamäärän omaava solmu 282 eli yli 1,2 miljoonaa kertaa katsottu video ei pääse lainkaan rakenteellisten keskeisyysmittareiden perusteella valittujen joukkoon. Tämän solmun asteeksi huomataan 0, ja tarkempi tutkimus osoittaa, että videon kommentointi on poistettu käytöstä tältä videolta. Toiseksi katsotuinta videota (solmu 262) on kuitenkin kommentoitu, eikä se silti pääse rakenteellisten keskeisyysmittareiden perusteella kärkikymmenikköön. Vain solmut 201 ja 183 esiintyvät katselumäärän lisäksi muissa keskeisyysarvosarakkeissa.

Taulukko 6. Solmujen pisteytys (10 keskeisintä graafissa 2).

Solmu	Aste	Painotettu aste	Ominais-vektori	Läheisyys	Välillisyys	Katselu-kerrat	Pisteet yhteensä
125	10	9	9	10	9		47
130	10	10	10	8			38
209	9	6	6	9			30
35	8	5	8	2			23
183	7	5	5	4		2	23
119	7	5		7			19
112	7			3	8		18
160	9	8					17
177	7		4	6			17
235	8	5			4		17

Myös suurimman välillisyyden omaava solmu 204 puuttuu suurimpien kokonaiskeskeisyyspisteiden solmujoukosta. Solmulla voi siis olla korkea välillisyys, vaikka se muilla mittareilla arvioituna ei yltäisi keskeisimpien solmujen joukkoon.

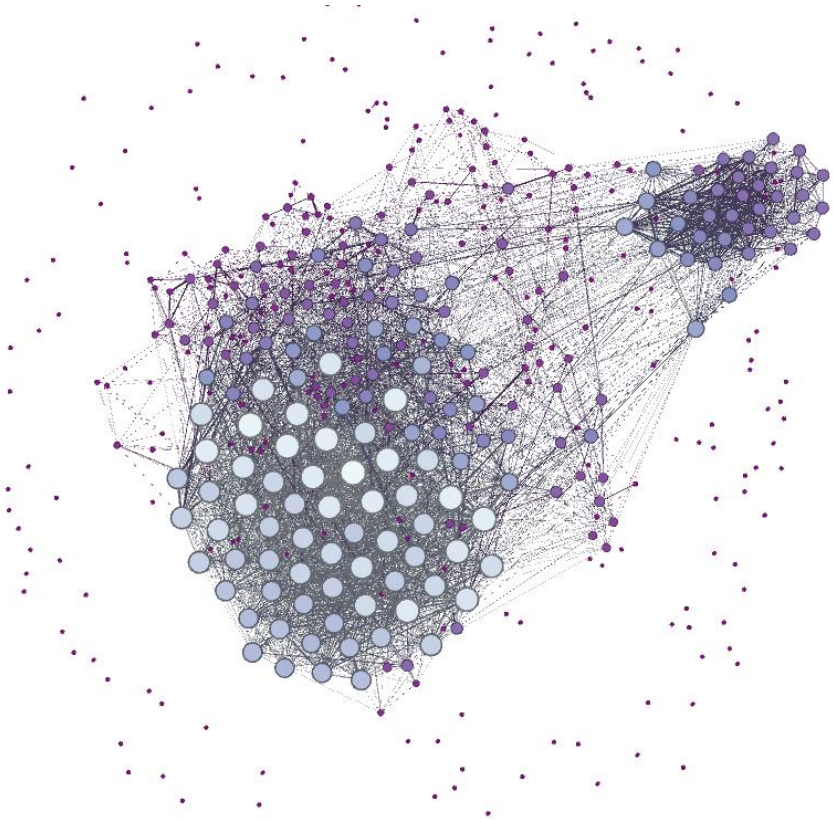


Kuva 22. Suurimmat kokonaiskeskeisyyspisteet saaneet solmut graafissa 2.

Kuvassa 22 esitetään suurimmat kokonaiskeskeisyyspisteet saaneet solmut. Katselumäärältään suurimmista videoista vain yksi (solmu 183) pääsee suurimpien kokonaiskeskeisyyspisteiden mukaan järjestettyyn joukkoon. Solmut 125 ja 130 erottautuvat muista suurimpien kokonaiskeskeisyyspisteiden perusteella, mutta solmu 125 saa suuremmat kokonaispisteet korkean välillisyytensä ansiosta.

3. Videoverkosto 3 (”home organizing”)

Kuvassa 23 havainnollistetussa videoverkostossa 3 on 465 solmua ja 5076 särmää. Graafissa on 152 komponenttia, joista 149 on yksisolmuisia (32 % solmuista). Suurimmassa komponentissa on 312 solmua (67,1 % graafista). Särmän paino on suurimmillaan 7. Särmien suuresta lukumäärästä johtuen solmujen suurin aste on myös korkea: 96. Tämä ja muut suurimmat keskeisyysarvot löytyvät taulukosta 6. Graafin modulaarisuus on 0,36.



Kuva 23. Yleiskuva koko graafista 3 (”home organizing”).

Taulukosta 7 huomataan, että yksikään kymmenestä katselukerroiltaan suurimmasta solmusta ei esiinny muualla taulukossa. Samasta syystä taulukosta 8 havaitaan, että suurimmat kokonaiskeskeisyyspisteet saaneet solmut eivät ole katselumäärältään kymmenen suurimman joukossa. Myös tässä verkostossa

suurimman välillisyyden omaava solmu (53) ei ole muilla mittareilla arvioituna keskeisimpien joukossa.

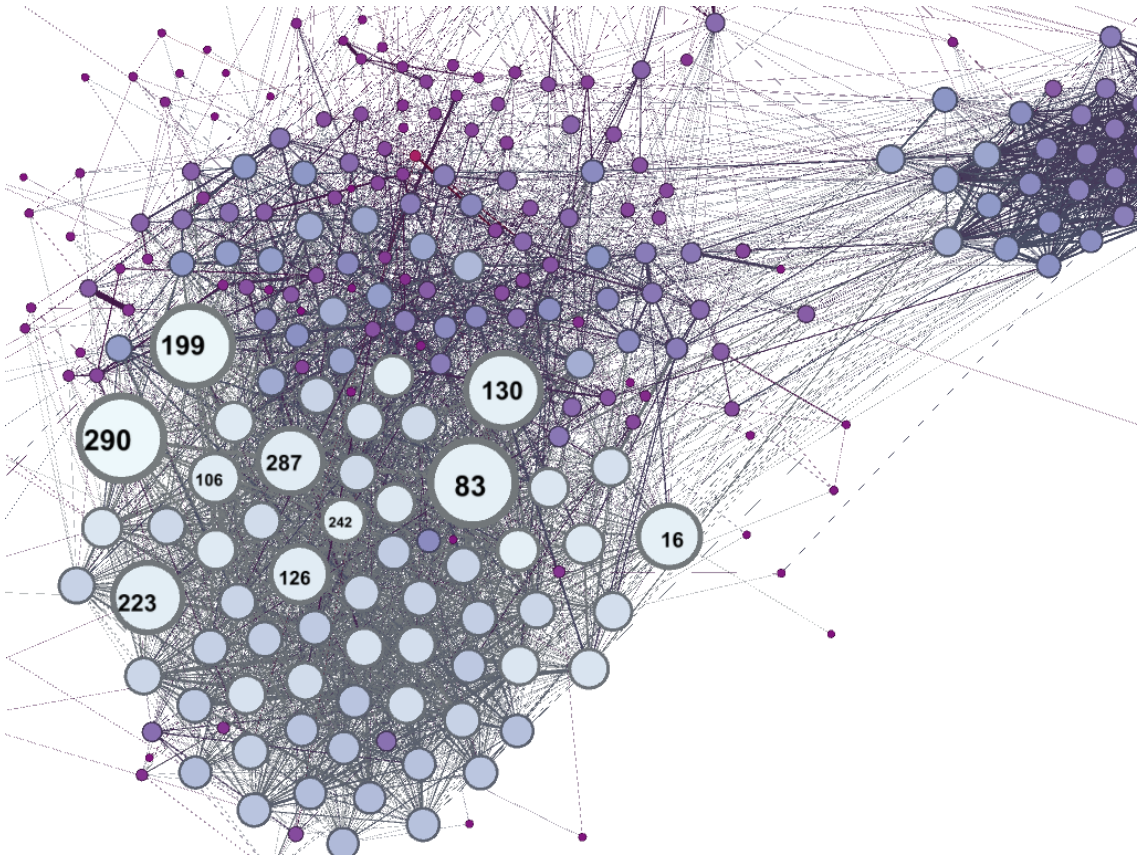
Taulukko 7. Keskeisyysarvoja (10–12 ”parasta” solmua graafissa 3).

Aste		Painot. aste		Ominaisvektori		Läheisyys		Välillisyys		Katsojamäärä	
arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu
96	290	155	83	1,000	290	1,826	199	1452,380	53	2007378	315
95	199	144	130	0,986	287	1,830	223	1401,792	223	1176809	232
92	16	142	287	0,967	106	1,836	83	1296,187	16	1044232	296
92	83	138	39	0,966	199	1,836	126	1266,590	70	898343	294
92	130	134	126	0,964	242	1,836	130	1135,326	42	793667	57
92	180	130	129	0,961	83	1,839	276	991,271	39	695055	309
92	242	128	290	0,960	180	1,839	290	962,170	15	645691	288
91	223	128	183	0,960	275	1,849	16	946,509	165	565248	284
91	275	127	106	0,953	130	1,849	35	864,780	51	560323	206
90	106	126	199	0,953	286	1,855	129	813,719	199	506484	312
90	126	126	286			1,855	291				
90	287										

Taulukko 8. Solmujen pisteytys (10 keskeisintä graafissa 3).

Solmu	Aste	Painotettu aste	Ominais-vektori	Läheisyys	Välillisyys	Katselu-kerrat	Pisteet yhteensä
83	8	10	5	8			31
290	10	4	10	7			31
199	9	2	7	10	1		29
130	8	9	2	8			27
223	7			9	9		25
287	6	8	9				23
16	8			6	8		22
126	6	6		8			20
106	6	3	8				17
242	8		6				14

Kuvassa 24 esitetään suurimmat kokonaiskeskeisyyspisteet saaneet solmut. Kuvan selkeyden parantamiseksi graafi on kuvassa vain osittain. Punaisella merkitty pieni solmu on 315, jolla on suurin katsojamäärä. Tässä videoverkostossa solmut 83 ja 290 saavat korkeimman pistearvon 31.

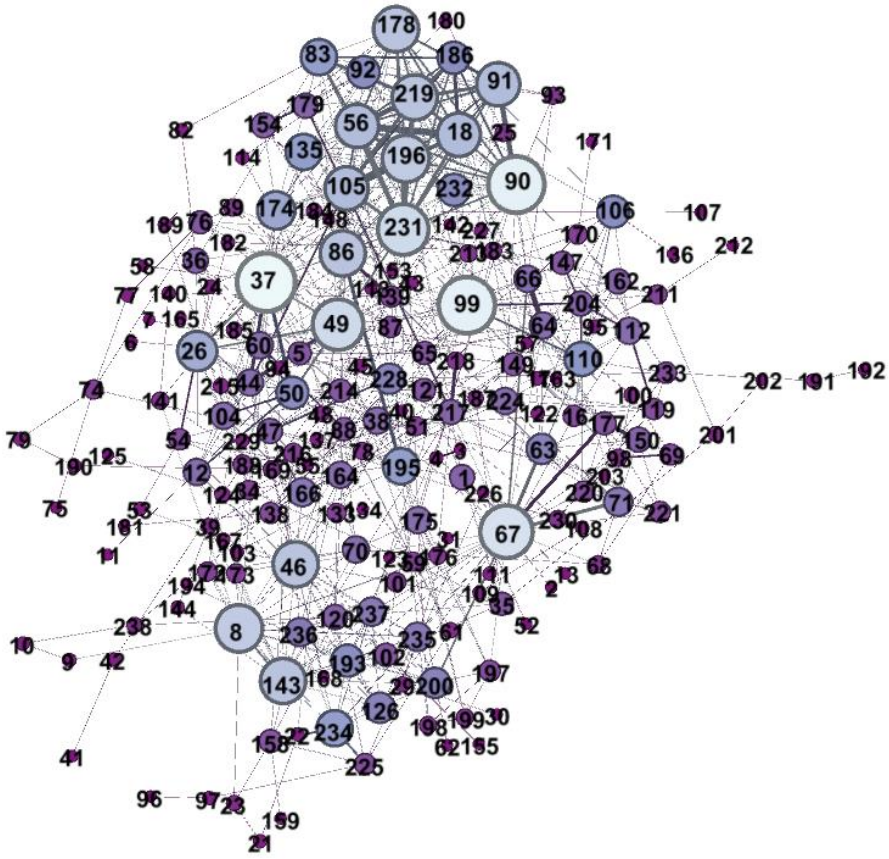


Kuva 24. Suurimmat kokonaiskeskeisyyspisteet saaneet solmut graafissa 3.

4. Videoverkosto 4 (50's pin up rockabilly makeup tutorial)

Videoverkostossa 4 on 463 solmua ja 641 särmää. Yleiskuva koko graafista esitetään kuvassa 25. Komponentteja on 243, joista suurimmassa on 198 solmua (42,76 % graafista). Tämä esitetään kuvassa 26. Yksisolmuisia komponentteja on 225 eli 48,6 % graafista. Koko graafin modulaarisuus on 0,598. Suurin solmun aste on 26, ja särmän paino on suurimmillaan 6.

Kuva 25. Yleiskuva koko graafista 4 (50’s pin up rockabilly makeup tutorial).



Kuva 26. Suurin komponentti graafissa 4.

Myös tässä verkostossa katselukerroiltaan suurimmat 10 videota eivät pääse muiden keskeisyysmittareiden perusteella 10–13 ensimmäisen solmun joukkoon, jotka esitetään taulukossa 9. Tästä johtuen katselukertojen sarake on tyhjillään myös taulukossa 10. Kaikkia katselukerroiltaan suurimpia videoita oli kuitenkin kommentoitu.

Taulukko 9. Keskeisyysarvoja (10–13 ”parasta” solmua graafissa 4).

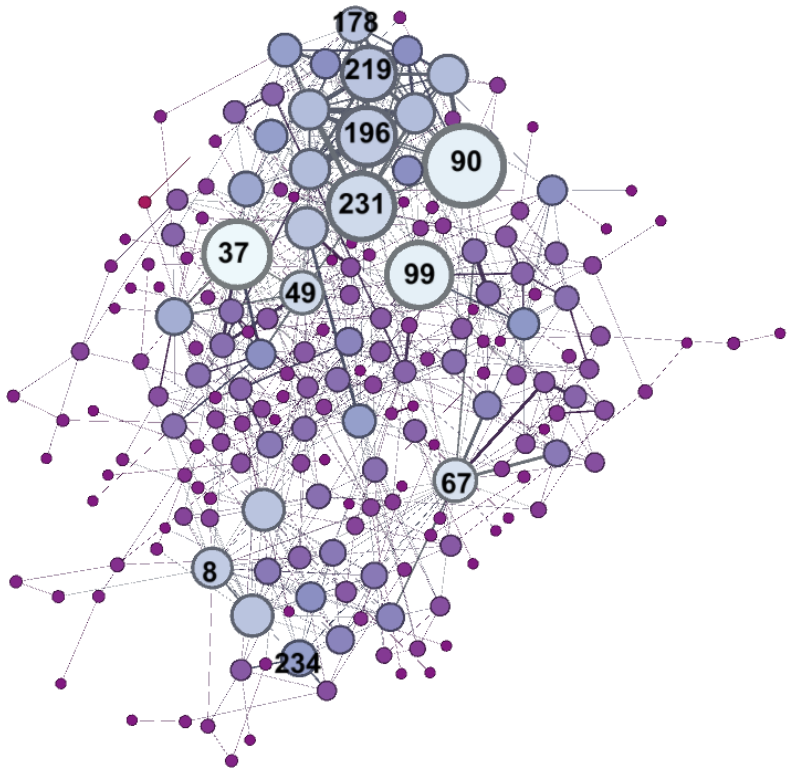
Aste		Painot. aste		Ominaisvektori		Läheisyys		Välillisyys		Katselukerrat	
arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu
26	37	40	231	1,000	90	2,462	90	2595,524	99	3449511	189
25	90	38	219	0,994	231	2,482	99	1809,440	49	2453331	166
25	99	37	196	0,984	37	2,523	37	1663,901	234	1426202	162
23	67	36	56	0,937	196	2,629	231	1555,967	90	887075	224
22	49	35	18	0,930	178	2,645	196	1383,155	8	842147	164
22	231	35	90	0,911	219	2,650	105	1368,158	46	836634	120
20	8	32	37	0,906	99	2,650	219	1334,950	67	766363	203
20	196	32	105	0,903	91	2,675	86	1303,359	143	647153	119
19	46	31	67	0,899	56	2,680	234	1259,107	37	581485	187
19	86	29	91	0,869	18	2,685	8	1054,819	174	558110	113
19	143										
19	178										
19	219										

Taulukossa 10 ja kuvassa 27 esitetään suurimmat kokonaiskeskeisyyspisteet saaneet solmut.

Taulukko 10. Solmujen pisteytys (11 keskeisintä graafissa 4).

Solmu	Aste	Painotettu aste	Ominais-vektori	Läheisyys	Välillisyys	Katselu-kerrat	Pisteet yhteensä
90	9	6	10	10	7		42
37	10	5	8	8	2		33
231	7	10	9	7			33
99	9		4	9	10		32
196	6	8	7	6			27
219	5	9	5	5			24
49	7				9		16
67	8	4			4		16
8	6			2	6		14
178	5		6				11
234				3	8		11

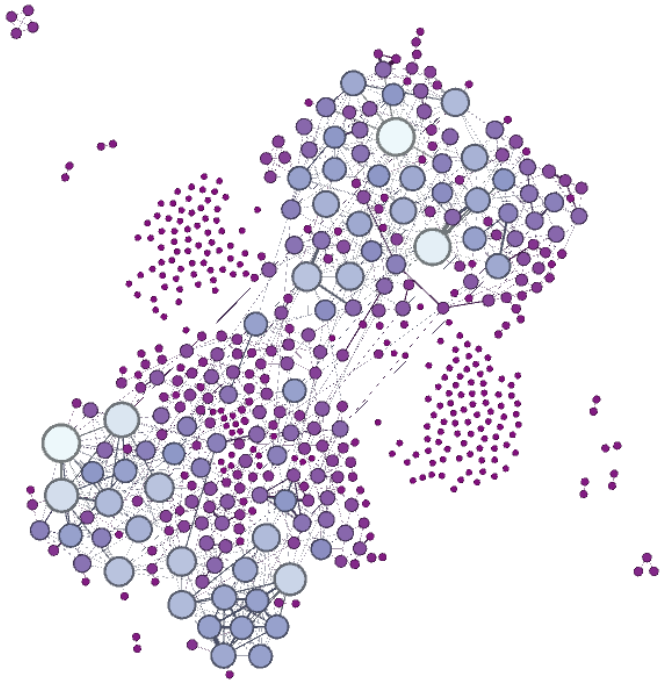
Tässä verkostossa solmu 234 on kokonaiskeskeisyyspisteiden perusteella 11. keskeisin solmu, vaikka se ei asteen tai painotetun asteen perusteella pääse keskeisimpien joukkoon. Tämä solmu on korkealla välillisyysarvonsa perusteella. Kuvasta 27 nähdään, että kyseinen solmu sijaitsee komponentin reuna-osassa, mutta siitä on kuitenkin vielä yhteyksiä komponentin uloimpiin solmuihin. Täten se on myös lukuisten geodeesisten polkujen varrella.



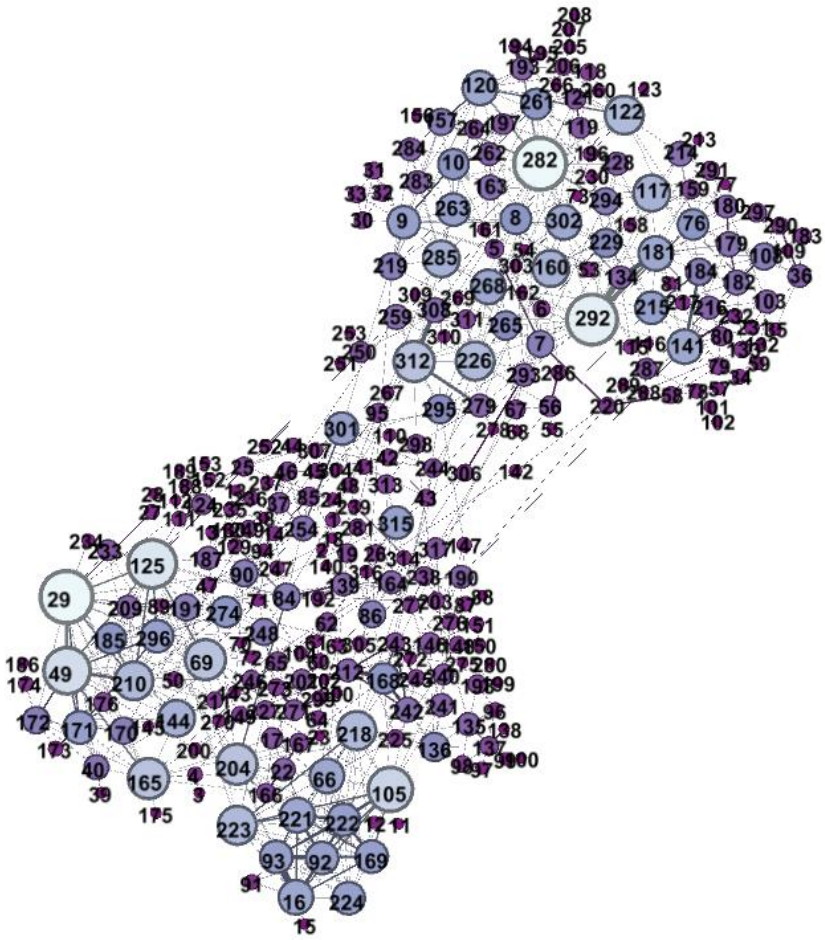
Kuva 27. Suurimmat kokonaiskeskeisyyspisteet saaneet solmut graafissa 4.

5. Videoverkosto 5 (How to cut your own hair)

Tämän tutkimuksen viimeisessä videoverkostossa on 500 solmua ja 821 särmää. Koko verkosto esitetään kuvassa 28. Graafi koostuu 194 komponenteista, joista suurimmassa on 294 solmua (58,8 % koko graafista). Suurin komponentti esitetään kuvassa 29. Irrallisia solmuja on 183 (36,6 %). Solmujen asteet ovat välillä 0–22 ja särmien painot välillä 1–6. Koko graafin modulaarisuus on 0,684.



Kuva 28. Yleiskuva koko graafista 5 (How to cut your own hair).



Kuva 29. Suurin komponentti graafissa 5.

Taulukosta 11 huomataan, että myös tässä videoverkostossa katselukerroiltaan suurimmat solmut eivät yhtä solmua (295) lukuun ottamatta pääse muiden keskeisyysmittareiden perusteella keskeisimpien 10–11 ensimmäisen solmun joukkoon. Täten myös tässä verkostossa katselukertojen sarake on tyhjä keskeisyyspisteiden joukossa, kuten taulukosta 12 havaitaan.

Taulukko 11. Keskeisyysarvoja (10–11 ”parasta” solmua graafissa 5).

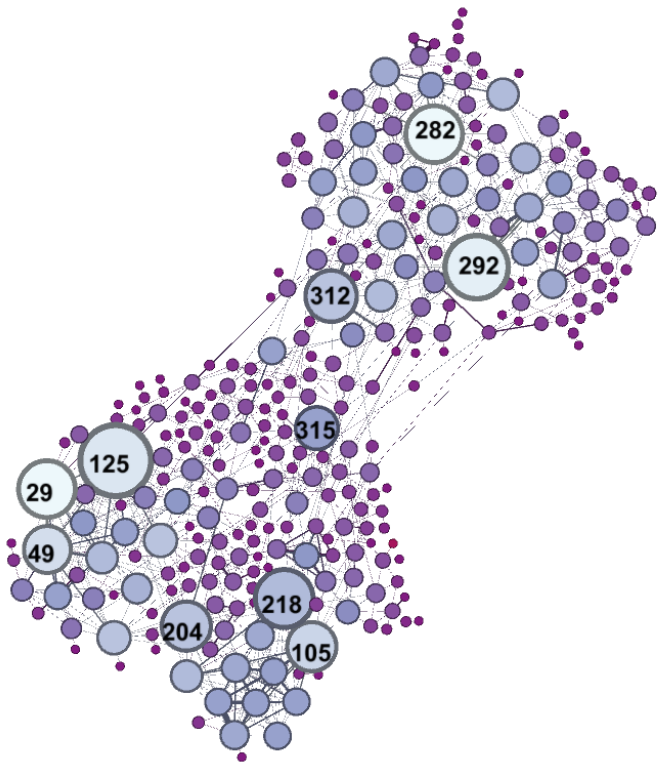
Aste		Painot. aste		Ominaisvektori		Läheisyys		Välillisyys		Katselukerrat	
arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu	arvo	solmu
22	29	28	49	1,000	204	3,246	312	4655,946	125	5375172	280
22	282	28	292	0,976	223	3,246	218	3757,351	218	3483508	238
21	292	27	29	0,943	218	3,259	125	3449,265	282	3079271	332
20	125	26	282	0,940	105	3,287	226	3007,050	315	2439742	326
19	49	25	125	0,909	29	3,294	292	3004,319	292	2073172	87
18	105	24	16	0,890	125	3,304	315	2907,855	301	1821287	246
16	69	24	93	0,863	221	3,334	268	2787,575	250	1805724	100
16	165	24	105	0,839	16	3,338	204	2611,912	190	1633955	55
16	204	22	92	0,837	222	3,358	295	2565,496	312	1607475	107
16	312	21	221	0,834	66	3,372	301	2470,203	29	1557374	295
		21	312								

Taulukko 12. Solmujen pisteytys (10 keskeisintä graafissa 5).

Solmu	Aste	Painotettu aste	Ominais-vektori	Läheisyys	Välillisyys	Katselu-kerrat	Pisteet yhteensä
125	8	7	5	9	10		39
292	9	10		7	6		32
218			8	10	9		27
29	10	9	6		1		26
282	10	8			8		26
312	5	4		10	2		21
105	6	6	7				19
204	5		10	4			19
49	7	10					17
315				6	7		13

Solmu 218 on siitä mielenkiintoinen, että se on kokonaiskeskeisyyspisteissä kolmanneksi suurimpana, vaikka ei yllä asteen ja painotetun asteen perusteella pistesijoille. Se on kuitenkin lähellä muita komponentin solmuja ja vieläpä

”tärkeiden” solmujen lähellä. Solmun 218 ja muut kokonaiskeskeisyyspisteiden perusteella merkittävät solmut voi paikallistaa kuvasta 30.



Kuva 30. Suurimmat kokonaiskeskeisyyspisteet saaneet solmut graafissa 5.

Solmut järjestettiin paremmuusjärjestykseen myös skaalaamalla keskeisyyspisteet kaikkien solmujen kesken verkostoittain ja laskemalla skaalatut pisteet yhteen. Tämän laskutavan mukaan keskeisimmät 10 solmua eivät juurikaan eronneet aiemmin esitellyistä solmuista. Eri laskutavoilla mitattuna oli pääosin vain yksittäisiä solmuja koskevia eroja, ja keskeisimmät solmut saattoivat olla eri järjestyksessä.

9.4. Keskeisyysarvojen ja attribuuttidatan väliset riippuvuudet

Rakenteellisille keskeisyysarvoille ja videoiden attribuuttidatalle laskettiin Pearsonin korrelaatiokertoimet videoverkostoittain. Näistä saatujen keskiarvojen mukaan keskeisyysarvojen ja attribuuttidatan välillä on vain heikkoa, lineaarista riippuvuutta. Taulukossa 13 esitellään keskiarvot eri keskeisyysarvojen

ja katselukertojen korrelaatiokertoimista. Vain välillisyyden ja katselukertojen välinen korrelaatio on hieman yli 0,3, mutta muut korrelaatiot ovat alle 0,3.

Taulukko 13. Keskeisyysarvojen ja katselukertojen korrelaatiokertoimien keskiarvot.

		r_aste_ katselukerrat	r_painotettu_aste_ katselukerrat	r_ominaisvektori_ katselukerrat	r_läheisyys_ katselukerrat	r_välillisyys_ katselukerrat
N	Valid	5	5	5	5	5
	Missing	0	0	0	0	0
Mean		,24950	,25671	,19904	,22955	,30251

Taulukossa 14 kuvataan keskiarvot eri keskeisyysarvojen ja arvosanojen korrelaatiokertoimista. Tuloksista huomataan, että riippuvuus arvosanojen ja keskeisyysarvojen välillä on vielä heikompaa kuin katselukertojen ja keskeisyysarvojen korrelaatio.

Taulukko 14. Keskeisyysarvojen ja arvosanojen korrelaatiokertoimien keskiarvot.

		r_aste_arvosana	r_painotettu_aste_ arvosana	r_ominaisvektori_ arvosana	r_läheisyys_ arvosana	r_välillisyys_ arvosana
N	Valid	5	5	5	5	5
	Missing	0	0	0	0	0
Mean		,11298	,10780	,09263	,08720	,06437

Pearsonin korrelaatiokertoimet laskettiin myös eri keskeisyysarvojen välillä. Taulukossa 15 esitellään asteen ja muiden keskeisyysarvojen korrelaatiokertoimista lasketut keskiarvot. Asteen ja painotetun asteen välinen korrelaatio on hyvin voimakasta: yli 0,9. Samoin asteen ja ominaisvektorin välinen korrelaatio on yli 0,9. Myös asteen ja välillisyyden välinen riippuvuus on voimakasta: yli 0,7. Asteen ja läheisyyden välillä sen sijaan on vain matalaa riippuvuutta: alle 0,5.

Taulukko 15. Asteen ja muiden keskeisyysarvojen korrelaatiokertoimien keskiarvot.

		r_aste_ painotettu_ aste	r_aste_ ominaisvektori	r_aste_ läheisyys	r_aste_ välillisyys
N	Valid	5	5	5	5
	Missing	0	0	0	0
Mean		,97940	,94020	,48440	,73840

Taulukossa 16 esitellään muiden tässä tutkimuksessa laskettujen keskeisyysarvojen välisiä korrelaatiokertoimien keskiarvoja. Myös painotetun asteen ja ominaisvektorin välinen riippuvuus on yli 0,9 eli hyvin voimakasta. Samoin

painotetun asteen ja välillisyyden välinen korrelaatio on voimakasta: yli 0,7. Painotetun asteen ja ominaisvektorin korrelaatio läheisyyden kanssa on melko matalaa: noin 0,4. Läheisyyden ja välillisyyden välillä on vain heikkoa riippuvuutta: alle 0,3.

Taulukko 16. Keskeisyysarvojen korrelaatiokertoimien keskiarvoja.

		r_painotettu_aste_ominaisvektori	r_painotettu_aste_läheisyys	r_painotettu_aste_välillisyys	r_ominaisvektori_läheisyys	r_ominaisvektori_välillisyys	r_läheisyys_välillisyys
N	Valid	5	5	5	5	5	5
	Missing	0	0	0	0	0	0
Mean		,93500	,44720	,72280	,37660	,62200	,28960

Korrelaatiokertoimet laskettiin myös videoiden attribuuttidatan kesken. Näiden korrelaatiokertoimien keskiarvot esitellään taulukossa 17. Katselukertojen ja arvosanan välillä on hyvin heikkoa riippuvuutta, alle 0,1. Sen sijaan katselukertojen ja kommenttien lukumäärän välinen riippuvuus on voimakasta, yli 0,7. Vaikuttaa luonnolliselta, että kommenttien määrä kasvaa katselukertojen lisääntyessä.

Taulukko 17. Attribuuttidatan korrelaatiokertoimien keskiarvot.

		r_katselukerrat_arvosana	r_katselukerrat_kommentit
N	Valid	5	5
	Missing	0	0
Mean		,05320	,71540

10. Pohdinta

10.1. Graafien rakenteen pohdinta

Vaikka kaikki viisi videoverkostoa ovat kooltaan erilaisia, niistä löytyy rakenteeltaan samanlaisia elementtejä. Mislove ja muut [2007] toteavat, että internetin sosiaalisten verkostojen suurin, heikosti yhtenäinen komponentti on rakenteellisesti verkoston mielenkiintoisin osa. Heidän mukaansa kyseiseen komponenttiin kuulumattomat solmut ovat yleensä osa hyvin pieniä, eristäytyneitä klustereita tai eivät ole lainkaan yhteydessä muiden solmujen kanssa. Nämä havainnot vastaavat tämän tutkimuksen verkostoja, joissa on kaikissa yksi suhteellisen iso komponentti muihin graafin komponentteihin verrattuna. Tämän tutkimuksen komponentit tosin eivät ole heikosti yhtenäisiä, koska graafit ovat suuntaamattomia. Ensimmäinen videoverkosto poikkeaa kuitenkin muista, koska sen suurinkin komponentti on melko pieni. Yhden suuremman komponentin lisäksi kaikkia verkostoja yhdistää suuri yksittäisten komponenttien eli irrallisten solmujen osuus. Vaikka tässä tutkimuksessa videodataa etsittiin yhteisten kommentoijien muodostamien yhteyksien perusteella, irralliset solmut osoittavat, että on paljon videoita, joilla ei ole yhteistä kommentoijaa. Syy saattaa löytyä kommentoinnin kieltämisestä, vähäisestä kommentointimäärästä tai yksinkertaisesti siitä, että samat katsojat kommentoijat esimerkiksi vain tietyn lataajan videoita. Tosin tulee muistaa, että videoverkostot ovat tässä tutkimuksessa vain otoksia kokonaisista tietyn aiheen mukaisista videoverkostoista, joissa saattaa olla satojatuhansia videoita. Lisäksi verkostojen dynaamisuus vaikuttaa jatkuvasti uusien yhteyksien syntyyn.

Mislove ja muut [2007] totesivat tutkimustensa perusteella, että internetin sosiaalisissa verkostoissa on suuri, voimakkaasti toisiinsa yhteydessä oleva solmujen ydin, jota ympäröi paljon pienempiä klustereita. Ydinalueen solmuilla on täten korkea aste, ja ympäröivillä solmuilla on pieni aste. Tämän tutkimuksen verkostot vastaavat kyseistä päätelmää.

Tämän tutkimuksen videoverkostojen modulaarisuus vaihtelee välillä 0,133–0,684 ja on keskimäärin 0,472. Fortunato ja Castellano [2007] huomauttavat, että modulaarisuusarvoja ei voi vertailla graafien kesken, koska graafin koko vaikuttaa arvoihin. Täten tämän tutkimuksen verkosto, jolla on suurin modulaarisuus, ei välttämättä ole ”paras” yhteisörakenteeltaan. Joka tapauksessa tässä tutkimuksessa kolmen verkoston modulaarisuus on yli 0,5 ja yhden yli 0,3, jo-

ten yhteisöjen rakenteet ovat myös modulaarisuuden perusteella huomionarvoisia. Vain Paulo Coelhon kirja-aiheisen videoverkoston modulaarisuus on 0,133, mikä viittaa heikompaan yhteisörakenteeseen. Tähän todennäköisesti kuitenkin vaikuttaa suuri irrallisten solmujen osuus graafissa.

10.2. Yhteisöllisyyden pohdinta

Vaikka tämän tutkimuksen verkostojen solmut edustavat videoita, taustalla toimivat käyttäjät. Tutkimuksen videoverkostoja ei olisi olemassa ilman käyttäjiä, koska videoiden yhteydet muodostuvat yhteisten kommentoijien perusteella. Täten videoverkostoissa muodostuu myös sosiaalisia yhteisöjä. Vain hakusanoilla ”New York travel tips” haettu videoverkosto ei vaikuttanut sisältävän yhteisöjä yhtä solmuparia lukuun ottamatta. Paulo Coelhon kirja-aiheisesta verkostosta löytyi melko pieni komponentti, mutta se on kuitenkin selkeästi melko tiivis yhteisö, jossa jopa kaikki keskeisimmät solmut olivat toistensa kanssa yhteydessä.

YouTuben yhteisö voidaan määritellä tiettyyn kategoriaan kuuluvien videoiden ryhmäksi [Susarla *et al.*, 2012]. Tämän tutkimuksen videoverkostojen kategorioita ei tutkittu tarkemmin, mutta videoiden voidaan katsoa kuuluvan saman ryhmään yhteisen aiheen perusteella. Graafeista kuitenkin huomattiin, että jokaisessa videoverkostossa esiintyi irrallisia solmuja, joilla ei ollut yhteyksiä muihin solmuihin. Koska yhteisön määritelmään kuuluu graafin yhtenäisyys, nämä irralliset solmut eivät siten kuulu tämän tutkimuksen yhteisöihin. Tämä ei kuitenkaan tarkoita, etteivätkö kyseiset solmut voisi kuulua joihinkin muihin yhteisöihin. Esimerkiksi jos yhteisöt olisi tässä määritelty pelkästään videoiden aiheen perusteella, eikä yhteisiä kommentoijia olisi otettu huomioon, kaikki haetut yhden videoverkoston videot olisivat kuuluneet samaan yhteisöön. Yhteisöjä voisi tarvittaessa osittaa pienimmäksi klustereiksi jonkin ominaisuuden perusteella.

10.3. Keskeisyystulosten pohdinta

Videoverkostoista laskettujen solmujen keskeisyysarvojen ja pisteytysten perusteella vaikuttaa siltä, että videoiden katselukerrat ja rakenteelliset keskeisyysmittarit eivät ole toisistaan riippuvaisia. Päätelmää tukevat myös matalat korrelaatiokertoimet rakenteellisten keskeisyysarvojen ja katselukertojen välillä. Kolmessa videoverkostossa kymmenen katsotuinta videota eivät esiintyneet muiden keskeisyysarvojen kärkijoukossa ollenkaan. Ainoastaan ensimmäisessä videoverkostossa katselumäärältään suurin video oli myös muiden keskeisyysmittareiden mukaan ensimmäisenä. Ensimmäinen videoverkosto oli kui-

tenkin muista poikkeava, koska sen suurin komponentti sisälsi vain 19 solmua ja loput komponenteista olivat yksisolmuisia. Täten myös parhaat keskeisyysarvot vaikuttivat jakautuvan helpommin samojen solmujen kesken.

Ei voida sanoa, että rakenteellisten keskeisyysmittareiden perusteella keskeisimmiksi valitut videot olisivat ”tärkeämpiä” kuin katsojamäärältään suurimmat. On huomioitava, että jos videosta on poistettu kommentointi käytöstä, se ei voi saada keskeisyysmittareiden perusteella pisteitä, vaikka katselumäärä olisi videoiden suurin. Toisaalta videoiden katselumäärät eivät välttämättä kuvaa todellista keskeisyyttä ja videoiden sosiaalista vaikutusvaltaa. Katselumäärien perusteella ei voida arvioida, onko videoita katsottu kokonaan. Videoiden suuret katsojamäärät saattavat myös olla tuloksia kanavien huomionhakuisuusyrityksistä, joilla pyritään houkuttelemaan videoille ja kanaville uusia katsojia [Susarla *et al.*, 2012].

Suuria katselumääriä etsimällä voidaan joka tapauksessa löytää videoita, jotka saavuttavat suuren yleisön huomion. On kuitenkin huomattava, että YouTuben videoverkostosta voidaan löytää myös pienempiä yhteisöjä, joista silti löytyy vaikutusvaltaa. Tällaisten yhteisöjen ja keskeisten videoiden löytämisessä voidaan hyödyntää keskeisyysmittareita ja tässä tutkielmassa esiteltyä kokonaiskeskeisyyspistemäärää. Myös Rotman ja Golbeck [2011] huomauttavat, että mainostajille voisi olla hyödyllistä etsiä yhteisöjä, joissa keskeiset videot ja käyttäjät eivät ehkä ole kaikkein suosituimpia, mutta kuitenkin vaikutusvaltaisia omassa aliyhteisössään. Mainostajien lisäksi tällaisten yhteisöjen keskeisimpien videoiden löytämistä voisi hyödyntää esimerkiksi suosittelu- ja hakutoiminnoissa, jotka auttavat YouTuben käyttäjiä löytämään videoita.

Vaikka katselukertojen perusteella merkittävimmät videot eivät näyttäneet pääsevän kokonaispisteiden perusteella keskeisimpien joukkoon, on keskeisyyspisteiden laskemisesta hyötyä. Keskeisyyspisteiden yhteenlaskun avulla solmujen keskeisyyttä voidaan arvioida kaikkien keskeisyysmittareiden perusteella, mikä tuo lisäarvoa verrattuna yhden keskeisyysmittarin perusteella tehtäviin arvioihin. Kokonaiskeskeisyyspisteitä voidaan myös käyttää hyödyksi silloin, kun verkoston yhteyksien luonne on sellainen, että keskeisimpiä solmuja ei välttämättä löydetä vain yhtä keskeisyysmittaria käyttämällä. Esimerkiksi läheisyysmittarin perusteella keskeisimmät solmut saattavat kuulua hyvin pieniin, vaikkapa vain kahden solmun yhteisöihin. Tällaiset solmut eivät käytännössä liene koko verkoston kannalta kaikkein keskeisimpiä solmuja.

Tämän tutkimuksen rakenteellisten keskeisyysarvojen väliset korrelaatiot vaihtelivat heikosta (alle 0,3) hyvin voimakkaaseen (yli 0,9). Asteen, painotetun asteen ja ominaisvektorin väliset riippuvuudet olivat kaikki yli 0,9. Myös Valenten ja muiden [2008] tutkimuksissa keskeisyysarvojen välisten korrelaatioiden voimakkuus vaihteli ja suurimmillaan se oli asteen ja ominaisvektorin välillä (0,92). Heidän tutkimuksissaan toiseksi suurin korrelaatio havaittiin asteen ja välillisyyden välillä (0,85), joka on melko lähellä tämän tutkimuksen vastaavien mittareiden välistä korrelaatiota (noin 0,7). Kovin voimakas korrelaatio keskeisyysarvojen välillä voi merkitä, että keskeisyysmittarit ovat jokseenkin redundantteja [Valente *et al.*, 2008]. Toisaalta matala korrelaatio saattaa viitata siihen, että mittarit kuvaavat hyvin erilaisia topologisia ominaisuuksia. Tässä tutkimuksessa lasketuista keskeisyyspisteistä huomattiin, että solmujen ”paremmuusjärjestys” vaihteli myös voimakkaasti korreloivien keskeisyysarvojen välillä. Esimerkiksi kaikki asteen perusteella keskeisimmät solmut eivät olleet painotetun asteen tai ominaisvektorin perusteella keskeisimpien joukossa. Täten eri mittarit eivät olleet redundantteja ja niiden yhdistäminen kokonaiskeskeisyyspisteiksi oli hyödyllistä keskeisimpien solmujen löytämiseksi.

10.4. Jatkokehitys

Tässä tutkimuksessa ei huomioida YouTube-datan dynaamisuutta ja videoiden ikää. Yhtä videota koskeva data saattaa nimittäin muuttua huomattavasti ja hyvin nopeasti – uusia kommentteja tulee lisää, niihin vastataan ja vanhoja kommentteja ehkä poistetaan. Lisäksi videoiden katselukertojen lukumäärät ja arvioinnit ovat muuttuvaa dataa. Tätä tutkimusta voisi jatkaa esimerkiksi analysoimalla eri ajanjaksojen videodataa ja tutkia muutosten vaikutusta keskeisyyslukuihin.

Sosiaalisen median sovellusten datan dynaamisuuden lisäksi myös itse sovellukset muuttuvat ja kehittyvät jatkuvasti, mikä tuo haasteita tutkimukseen. Tämä näkyy myös monissa YouTuben ominaisuuksissa. Myös YouTuben kommentointitoimintoa on päivitetty ja kommentointi on yhdistetty Google+-palveluun. Päivityksen myötä käyttäjät voivat ilmaista tykkäävänsä myös kommenteista. Uudenlaiset ja muuttuvat toiminnot tuovat myös esille uusia mahdollisia tutkimuskohteita, joita tässä tutkimuksessa ei ole huomioitu.

Tässä tutkimuksessa esiteltyjen keskeisyyspisteiden ja skaalattujen keskeisyyspisteiden perusteella laskettujen keskeisimpien solmujen välillä oli hieman eroa videoverkostoittain. Esille nousseiden erojen syitä voisi selvittää tarkemmin, ja keskeisyyspisteiden laskentaa voisi testata myös muulla datalla. Tämän tutki-

muksen data kuvattiin suuntaamattomien graafien avulla. Keskeisyyspisteiden laskemista voisi kehittää ja soveltaa myös suunnatuille graafeille.

Lisäksi tutkimusta voisi jatkaa analysoimalla tarkemmin videoiden aiheen vaikutusta yhteisöllisyyteen ja keskeisyyslukuihin. Esimerkiksi tässä tutkimuksessa hakusanoilla ”New York travel tips” ei vaikuttanut löytyvän yhteisöllisyyttä haetusta videodatasta, koska 98 solmun joukossa oli vain yksi särmä. Tällainen matkailuun liittyvä aihe saattaa houkutella sellaisia satunnaisia katsojia ja kommentoijia, jotka eivät kuitenkaan seuraa ja kommentoi samanaiheisia videoita sen enempää. Lisätutkimusta suuremmalla datamäärällä kuitenkin tarvittaisiin, jotta sattuman vaikutusta voitaisiin arvioida.

Tutkimuksessa ei myöskään huomioitu videoihin liittyvää käyttäjäverkostoa. Tietyt käyttäjät saattavat nimittäin kommentoida vain joidenkin seuraamiensa kanavien videoita, jolloin myös näiden välille luonnollisesti syntyy enemmän yhteyksiä. Olisi myös mielenkiintoista tutkia, paljonko videoiden lataajat kommentoivat muiden lataajien videoita, ja minkälaista keskustelua lataajien välille mahdollisesti syntyy. Käyttäjäverkostoja voidaan tarkastella videoiden lataajien, videokanavien tilaajien, katsojien ja kommentoijien näkökulmista, joten käyttäjistä muodostuvat limittyvät yhteisöt voisi myös olla eräs tutkimuskohde.

Tämän tutkimuksen videoverkostojen särmät muodostuivat videoita yhdistävistä kommentoijista. Kommenttien sisältöä ja laatua ei kuitenkaan analysoitu tarkemmin. Olisi mielenkiintoista tutkia kommentteihin sisältyviä mielipiteitä ja kommenteissa esiintyviä keskusteluita. Kommenttidataa voisi analysoida *mielipiteiden louhinnan* (opinion mining) menetelmillä, joiden avulla mielipiteitä ja näkemyksiä voidaan poimia, luokitella, määritellä ja arvioida [Chen and Zimbra, 2010]. *Tunneanalyysi* (sentiment analysis) on eräs mielipiteiden louhinnassa käytettävä menetelmä, jonka perusteella kommentteihin sisältyviä ajatuksia, tunnereaktioita, kiihtymystä ja muita tunnetiloja voisi tunnistaa.

11. Yhteenveto

Sosiaaliset verkostot koostuvat ihmisten ja erilaisten asioiden välisistä yhteyksistä. Teknologian kehitys on tuonut sosiaalisten verkostojen analysointiin uuden näkökulman, koska virtuaaliset sosiaaliset verkostot ja sosiaalisen median sovellukset ovat yleistyneet ja monipuolistuneet. Suosituimmat sosiaaliset verkostot ovat myös kasvaneet räjähdysmäisesti, mikä luo omat haasteensa verkostojen tutkimiseen. Koska verkostot saattavat sisältää jopa miljardeja entiteettejä, tarvitaan keinoja löytää ja analysoida pienempiä kokonaisuuksia verkostoista. Verkostoja voidaan jakaa osiin eli klusteroida, ja verkostoista voidaan määritellä pienempiä yhteisöjä.

Verkostojen mallintamiseen ja analysointiin käytetään yleisesti formaalina lähestymistapana graafiteoriaa, jossa verkostojen entiteetit kuvataan solmuina ja entiteettien väliset yhteydet särminä. Yhteisö on ryhmä solmuja, joilla on jokin yhteinen ominaisuus tai tekijä. Verkostoissa saattaa olla irrallisia solmuja, joilla ei ole yhteyksiä muihin solmuihin. Verkostoista löytyvien yhteisöjen tulee kuitenkin olla yhtenäisiä, eli jokaisesta solmusta on polku yhteisön muihin solmuihin. Kahden solmun välille voidaan laskea geodeesinen etäisyys, joka on solmuja yhdistävä lyhin polku eli niiden välillä olevien särmien lukumäärä. Solmujen välinen geodeesinen etäisyys on merkittävä tekijä esimerkiksi solmujen läheisyys- ja välillisyyskeskeisyysarvoja laskettaessa.

Graafeja voidaan havainnollistaa visuaalisten esitystapojen ja erilaisten matriisien avulla. Esimerkiksi verkostojen kokoa, muotoa ja tiheyttä voidaan tutkia graafianalyysin menetelmillä. Lisäksi yksittäisten solmujen sijaintia ja solmujen ryhmittymiä voidaan tutkia ja paikallistaa graafista. Sosiaalisten verkostojen analysoinnissa tutkitaan erityisesti solmujen välisiä yhteyksiä ja etäisyyksiä.

Verkostoista voidaan löytää implisiittisiä yhteisöjä globaaleilla ja lokaaleilla lähestymistavoilla. Globaaleissa menetelmissä yhteisöjen etsiminen alkaa koko verkoston solmujen joukosta, mutta lokaaleissa etsintä voidaan aloittaa yhdestä solmusta. Globaalissa määrittelyssä yhteisöt ovat olennaisia osia koko verkoston toiminnalle. Globaaleissa menetelmissä verkoston jako yhteisöiksi määritellään yleensä yhteisöjen välisten särmien lukumääriä tarkastelemalla. Myös modulaarisuus voi olla globaali määrittelykriteeri. Lokaalisti määritellyt yhteisöt ovat yleensä maksimaalisia aligraafeja. Yhteisöjä voidaan löytää myös solmujen samankaltaisuuksien tai yhteisöjen muodostumisprosessien perusteella. Solmu-

jen samankaltaisuusmittoja käytetään esimerkiksi perinteisissä osittavissa ja hierarkkisissa klusterointimenetelmissä.

Osittavissa klusterointimenetelmissä graafi jaetaan osiin siten, että jokainen solmu kuuluu yhteen klusteriin. Käytännössä solmu voi kuitenkin kuulua useampaan eri ryhmään eli limittyviin yhteisöihin. Ositetut graafit voidaan myös järjestää hierarkkisesti, jolloin pienempi yhteisö kuuluu suurempaan yhteisöön. Graafien ositus on haastavaa, koska menetelmissä tulee löytää mahdollisimman optimaaliset ositukset lukuisista vaihtoehdoista. Ositusta voidaan evaluoida esimerkiksi modulaarisuuden mittarilla.

Hierarkkiset klusterointimenetelmät voidaan erotella jakaviin ja kokoaviin menetelmiin. Jakavissa menetelmissä verkosto jaetaan yhteisöihin poistamalla särmiä yksi kerrallaan. Eräs jakavista menetelmistä on Girvanin ja Newmanin kehittelemä särmien välillisyysarvoihin perustuva lähestymistapa. Kokoavissa menetelmissä solmut ryhmitellään suuremmiksi yhteisöiksi, kunnes koko verkosto on rakennettu. Solmujen yhdistämisessä käytetään solmujen välisiä läheisyysmittoja.

Yhteisöjä voidaan tarkastella kokonaisuuksina, esimerkiksi dynaamisen luonteen vuoksi niiden evoluutiota ja perustoimintoja voidaan tutkia. Yhteisöjä voidaan kuitenkin analysoida myös pienempinä kokonaisuuksina ja yksittäisten solmujen tasolla. Solmuille voidaan esimerkiksi määritellä erilaisia rooleja niiden rakenteellisen sijainnin perusteella. Solmun rooli voi kuvata solmun suhteita sen naapureihin ja laajempaan verkostorakenteeseen. Roolien perusteella voidaan esimerkiksi arvioida solmun vaikutusvaltaa yhteisössä tai määritellä solmuja, jotka toimivat siltoina eri yhteisöjen välillä.

Lisäksi solmujen keskeisyyttä voidaan arvioida erilaisten mittareiden näkökulmista. Solmun keskeisyys voi kuvata solmun tärkeyttä tai osallistumisen määrää yhteisössä. Yleisimmin käytetyt keskeisyysmittarit ovat aste, ominaisvektori, läheisyys ja välillisyys. Aste kuvaa solmun suorien yhteyksien lukumäärän, joten sen avulla voidaan arvioida solmun välitöntä vaikutusvaltaa tai riskiä verkostossa. Aste ei kuitenkaan mitenkään huomioi verkoston laajempaa rakennetta tai solmujen epäsuoria yhteyksiä muiden solmujen kautta. Asteesta voidaan laskea painotettu versio, jos graafin särmillä on painot. Ominaisvektorikeskeisyys suhteuttaa solmun keskeisyyden sen vierussolmujen keskeisyysarvoihin. Ominaisvektorikeskeisyyden avulla voidaan kuvata esimerkiksi vaikutusvaltaprosesseja, koska solmulla saattaa olla vaikutusvaltaa myös epäsuo-

rien yhteyksiensä kautta, jos se on ”tärkeän” solmun läheisyydessä. Läheisyys kuvaa, kuinka lähellä solmu on muita solmuja yhtenäisessä verkostossa. Läheisyyden avulla voidaan arvioida vaikkapa informaation kulkua, saavutettavuutta ja solmun itsenäisyyttä. Solmun välillisyys kuvaa, kuinka usein solmu osuu kaikkien solmuparien välisille geodeesisille poluille. Välillisyyden avulla voidaan arvioida solmun kontrollointimahdollisuuksia kommunikoinnin suhteen ja yhteisöjen välillä toimivien siltojen tärkeyttä verkoston yhteyksien kannalta.

Keskeisyysmittareihin liittyviä tutkimuksia on lukuisia, mutta useimmissa keskitytään mittareiden välisiin korrelaatioihin tai arvioidaan mittareiden ja verkoston rakenteellisten ominaisuuksien välisiä suhteita. Tässä tutkimuksessa solmujen keskeisyyttä arvioitiin yhdistämällä eri mittareilla laskettuja tuloksia. Tutkimuksessa käytettiin YouTubesta haettua videodataa, joka rajoitettiin kuuteen eri aihealueeseen. Videoita haettiin yhteensä 2271. Videoiden yhteystyyppinä on yhteinen kommentoija, joten löydettyjä yhteisöjä yhdistivät sekä videoiden yhteinen aihepiiri että yhteisten kommentoijien joukko. Vain hakusanoilla ”New York travel tips” löytyneistä 98 videosta ei muodostunut merkittävää yhteisöä, koska verkostossa oli vain yksi särmä. Muiden hakuaiheiden perusteella löytyi selkeää yhteisöllisyyttä, joten tutkimuksia jatkettiin viiden videoverkoston parissa.

Kaikki viisi videoverkostoa ovat erikokoisia, mutta niiden rakenteesta löytyy samankaltaisia elementtejä. Kaikissa verkostoissa on nimittäin yksi suhteellisen iso komponentti graafin muihin komponentteihin verrattuna. Nämä muut komponentit olivat myös hyvin pieniä tai irrallisia solmuja. Paulo Coelhon kirja-aiheinen videoverkosto erosi hieman muista, koska sen suurinkin komponentti oli melko pieni. Tästä yhtenäisestä komponentista löytyi kuitenkin keskeisimpien solmujen perusteella varsin tiivis, toistensa kanssa yhteydessä oleva yhteisö.

Jokaisen videoverkoston solmuille laskettiin keskeisyysarvot (aste, painotettu aste, ominaisvektori, läheisyys ja välillisyys) ja järjestettiin paremmuusjärjestykseen. Keskeisyystarkasteluun otettiin mukaan myös videoiden katselukerrat. Jokaisesta videoverkostosta tarkasteltiin noin kymmentä parasta solmua jokaisella eri mittarilla laskettuna ja näiden perusteella solmuille laskettiin keskeisyyspisteet. Lähempään tarkasteluun valittiin jälleen noin kymmenen keskeisintä solmua jokaisesta verkostosta kokonaiskeskeisyyspistemäärän perusteella. Lisäksi keskeisyysarvoille ja videoiden katselukerroille laskettiin korrelaatiokertoimet videoverkostoittain.

Solmujen keskeisyysarvojen ja -pisteiden perusteella huomattiin, että keskeisyysmittarit ja videoiden katselukerrat eivät ole toisistaan riippuvaisia. Tätä päätelmää tukevat myös matalat korrelaatiokertoimet keskeisyysarvojen ja katselukertojen välillä. Videoiden suuret katselukerrat kuvaavat videoita, jotka ovat saavuttaneet suuren yleisön huomion, mutta eivät välttämättä ole pienemmissä yhteisöissä keskeisiä ja vaikutusvaltaisia. Keskeisimpiä solmuja ei välttämättä löydä vain yhtä keskeisyysmittaria tai ominaisuutta tarkastelemalla. Tällöin voi olla hyödyllistä laskea solmuille kokonaiskeskeisyyspisteet tässä tutkimuksessa esitellyn pisteytyksen perusteella, jota voisi kehittää soveltumaan myös muunlaisille verkostoille.

Viiteluettelo

[Bastian *et al.*, 2009] Mathieu Bastian, Sebastien Heymann and Mathieu Jacomy, Gephi: an open source software for exploring and manipulating networks. In: *ICWSM'09: Proc. of the Third International AAAI Conference on Weblogs and Social Media* (Mar. 2009). Gephi available at: *https://gephi.org/*.

[Bodendorf and Kaiser, 2010] Freimut Bodendorf and Carolin Kaiser, Detecting opinion leaders and trends in online communities. In: *ICDS '10: Proc. of Fourth International Conference on Digital Society* (Feb. 2010), 124–129.

[Bonacich, 1972] Phillip Bonacich, Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* **2**, 1 (1972), 113–120. Also available at: *http://dx.doi.org/10.1080/0022250X.1972.9989806*. Checked 22.5.2014.

[Borgatti, 2005] Stephen P. Borgatti, Centrality and network flow. *Social Networks* **27**, 1 (2005), 55–71.

[Brandes, 2001] Ulrik Brandes, A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* **25**, 2 (2001), 163–177.

[Cai *et al.*, 2005] Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan and Jiawei Han, Community mining from multi-relational networks. In: *PKDD 2005: Proc. of Knowledge Discovery in Databases, Lecture Notes in Computer Science* **3721** (2005), Springer, 445–452.

[Cha *et al.*, 2007] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn and Sue Moon, I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: *IMC '07: Proc. of the 7th ACM SIGCOMM Conference on Internet Measurement* (Oct. 2007), 1–14.

[Chakrabarti, 2003] Soumen Chakrabarti, *Mining the Web – Discovering Knowledge from Hypertext Data*. Morgan Kaufman Publishers, USA, 2003.

[Chen and Zimbra, 2010] Hsinchun Chen and David Zimbra, AI and opinion mining. *IEEE Intelligent Systems* **25**, 3 (May-June 2010), 74–76.

[Chen *et al.*, 2009] Jiyang Chen, Osmar R. Zaïane and Randy Goebel, A visual data mining approach to find overlapping communities in networks. In: *ASONAM '09: Proc. of International Conf. on Advances in Social Network Analysis and Mining* (July 2009), 338–343.

[Cheng *et al.*, 2008] Xu Cheng, Cameron Dale and Jiangchuan Liu, Statistics and social network of YouTube videos. In: *IWQoS 2008: Proc. of the 16th International Workshop on Quality of Service* (June 2008), 229–238.

[Cheng *et al.*, 2013] Xu Cheng, Jiangchuan Liu and Cameron Dale, Understanding the characteristics of internet short video sharing: a YouTube-based measurement study. *IEEE Transactions on Multimedia* **15**, 5 (Aug. 2013), 1184–1194.

[Clauset, 2005] Aaron Clauset, Finding local community structure in networks. *Physical Review E* **72**, 2 (Aug. 2005). Available as *http://arxiv.org/pdf/physics/0503036v1.pdf*. Checked 6.3.2014.

[Csermely *et al.*, 2013] Peter Csermely, András London, Ling-Yun Wu and Brian Uzzi, Structure and dynamics of core/periphery networks. *Journal of Complex Networks*. Oxford University Press, 2013, 93–123.

[Duan *et al.*, 2009] Dongsheng Duan, Yuhua Li, Yanan Jin and Zhengding Lu, Community mining on dynamic weighted directed graphs. In: *CNIKM '09: Proc. of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management* (Nov. 2009), 11–18.

[Fortunato, 2010] Santo Fortunato, Community detection in graphs. *Physics Reports* **486**, 3–5 (Feb. 2010), 75–174. Available at: *http://arxiv.org/abs/0906.0612*. Checked 6.3.2014.

[Fortunato and Castellano, 2007] Santo Fortunato and Claudio Castellano, Community structure in graphs. In: R. A. Meyers (ed.), *Springer's Encyclopedia of Complexity and System Science*, 2008. Available at: *http://arxiv.org/abs/0712.2716*. Checked 6.3.2014.

[Freeman, 1979] Linton C. Freeman, Centrality in social networks – Conceptual clarification. *Social Networks* **1**, 3 (1978/1979), 215–239.

[Gephi Tutorial Layouts, 2014] Gephi tutorial layouts, website. *http://www.slideshare.net/gephi/gephi-tutorial-layouts*. Checked 26.4.2014.

[Girvan and Newman, 2002] Michelle Girvan and Mark E. J. Newman, Community structure in social and biological networks. *PNAS* **99**, 12 (June 2002), 7821–7826.

[Gleich and Seshadhri, 2012] David F. Gleich and C. Seshadhri, Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In: *Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2012), 597–605.

[Gould, 1967] P. R. Gould, On the geographical interpretation of eigenvalues. *Transactions of the Institute of British Geographers* 42 (Dec. 1967), 53–86.

[Han and Kamber, 2006] Jiawei Han and Micheline Kamber, *Data Mining – Concepts and Techniques*. Second Edition. Morgan Kaufman Publishers, USA, 2006.

[Hansen *et al.*, 2011] Derek L. Hansen, Ben Shneiderman and Marc A. Smith, *Analyzing Social Media Networks with NodeXL – Insights from a Connected World*. Morgan Kaufman Publishers, USA, 2011.

[Kang *et al.*, 2012] Chanhyun Kang, Cristian Molinaro, Sarit Kraus, Yuval Shavitt and V. S. Subrahmanian, Diffusion centrality in social networks. In: *ASONAM '12: Proc. of International Conf. on Advances in Social Network Analysis and Mining* (Aug. 2012), 558–564.

[Koivisto ja Niemistö, 2001] Pertti Koivisto ja Riitta Niemistö, Graafiteoriaa. Tampereen yliopisto, Kesäkuu 2001. Saatavilla: *http://www.uta.fi/sis/mattil/graafiteoria/graafiteoriaa.pdf*. Tarkistettu 6.3.2014.

[Kulkarni and Devetsikiotis, 2010] Vineet Kulkarni and Michael Devetsikiotis, Communication timescales, structure and popularity: using social network metrics for Youtube-like multimedia contend distribution. In: *ICC 2010: Proc. of the IEEE International Conference on Communications* (May 2010), 1–5.

[Landherr *et al.*, 2010] Andrea Landherr, Bettina Friedl and Julia Heidemann, A critical review of centrality measures in social networks. *Business & Information Systems Engineering* **2**, 6 (Dec. 2010), 371–385.

[Luo *et al.*, 2008] Feng Luo, James Z. Wang and Eric Promislow, Exploring local community structures in large networks. *Web Intelligence and Agent Systems: An International Journal* **6**, 4 (Dec. 2008), 387–400.

[Mislove *et al.*, 2007] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel and Bobby Bhattacharjee, Measurement and analysis of online social networks. In: *IMC '07: Proc. of the 7th ACM SIGCOMM Conference on Internet Measurement* (Oct. 2007), 29–42.

[Okamoto *et al.*, 2008] Kazuya Okamoto, Wei Chen and Xiang-Yang Li, Ranking of closeness centrality for large-scale social networks. In: *FAW 2008: Proc. of the Second Annual International Workshop* (June 2008), 186–195.

[Opsahl *et al.*, 2010] Tore Opsahl, Filip Agneessens and John Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* **32**, 3 (July 2010), 245–251.

[Palla *et al.*, 2005] Gergely Palla, Imre Derényi, Illés Farkas and Tamás Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** (June 2005), 814–818. Also available as *http://arxiv.org/pdf/physics.soc-ph/0506133.pdf*. Checked 14.5.2014.

[Papadopoulos *et al.*, 2011] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali and Ploutarchos Spyridonos, Community detection in social media – Performance and application considerations. *Data Min Knowl Disc*. **24**, 3 (Jan. 2012), 515–554.

[Radicchi *et al.*, 2004] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto and Domenico Parisi, Defining and identifying communities in networks. *PNAS* **101**, 9 (Mar. 2004), 2658–2663.

[Rotman and Golbeck, 2011] Dana Rotman and Jennifer Golbeck, YouTube – Contrasting patterns of content, interaction, and prominence. In: *Analyzing Social Media Networks with NodeXL – Insights from a Connected World*. Morgan Kaufman Publishers, USA, 2011, 225–246.

[Rotman *et al.*, 2009] Dana Rotman, Jennifer Golbeck and Jennifer Preece, The community is where the rapport is – on sense and structure in the YouTube community. In: *C & T '09*: *Proc. of the Fourth International Conference on Communities and Technologies* (June 2009), 41–50.

[Scott, 2000] John Scott, *Social Network Analysis, a Handbook*. Second Edition. SAGE Publications, 2000.

[Scripps *et al.*, 2007] Jerry Scripps, Pang-Ning Tan and Abdol-Hossein Esfahanian, Node roles and community structure in networks. In: *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (Aug. 2007), 26–35.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik, Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22,** 8 (Aug. 2000), 888–905.

[Smith *et al.*, 2010] Marc Smith, Natasa Milic-Frayling, Ben Shneiderman, Eduarda Mendes Rodrigues, Jure Leskovec and Cody Dunne, NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010. Available at: *http://nodexl.codeplex.com/* from the Social Media Research Foundation, *http://www.smrfoundation.org*. Checked 6.3.2014.

[Spathis and Gorcitz, 2011] Prométhée Spathis and Raul Adrian Gorcitz, A data-driven analysis of YouTube community features. In: *AINTEC '11 Proc. of the 7th Asian Internet Engineering Conference* (Nov. 2011), 12–18.

[Susarla *et al.*, 2012] Anjana Susarla, Jeong-Ha Oh and Yong Tan, Social networks and the diffusion of user-generated content: evidence from YouTube. *Information Systems Research* **23**, 1 (Mar. 2012), 23–41.

[Takaffoli *et al.*, 2011] Mansoureh Takaffoli, Farzad Sangi, Justin Fagnan and Osmar R. Zaïane, Community evolution mining in dynamic social networks. *Procedia – Social and Behavioral Sciences* **22** (2011), 49–58.

[Valente *et al.*, 2008] Thomas W. Valente, Kathryn Coronges, Cynthia Lakon and Elizabeth Costenbader, How correlated are network centrality measures? *Connect (Tor)* **28**, 1 (Jan. 2008), 16–26. Also available as

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875682/pdf/nihms80042.pdf. Checked 26.3.2014.

[Virtanen, 2003] Satu Virtanen, Properties of nonuniform random graph models. Helsinki University of Technology, Laboratory for Theoretical Computer Science, Research Reports **77**, 2003. Also available as *http://www.tcs.hut.fi/Publications/bibdb/HUT-TCS-A77.pdf*.

[Webb and Copsey, 2011] Andrew R. Webb and Keith D. Copsey, *Statistical Pattern Recognition*. Third Edition. John Wiley & Sons Ltd, 2011.

[Yang and Leskovec, 2012] Jaewon Yang and Jure Leskovec, Defining and evaluating network communities based on ground-truth. In: *ICDM 2012: Proc. of the 12th IEEE International Conference on Data Mining* (Aug. 2012). Also available as *http://cs.stanford.edu/people/jure/pubs/comscore-icdm12.pdf*. Checked 19.5.2014.

[YouTube, 2014] YouTube tilastotiedot, www-sivut. *https://www.youtube.com/yt/press/fi/statistics.html*. Tarkistettu 6.3.2014.

[YouTube API, 2014] YouTube Data API, Google Developers website. *https://developers.google.com/youtube/*. Checked 9.4.2014.