Bonferronin epäyhtälön käyttö assosiaatiosääntöjen yleistämisessä

Jouni K. Seppänen

Matematiikan pro gradu -työ
Helsingin yliopisto
Matematiikan laitos
Huhtikuu 2000

Sisältö

1 Johdanto

Tämä tutkielma käsittelee Bonferronin epäyhtälöä ja sen eräitä yleistyksiä. Yksinkertaisin kirjallisuudessa Bonferronin epäyhtälöksi kutsuttu tulos on lähes triviaali: joukkojen yhdisteen koko on enintään joukkojen kokojen summa,

$$\Big|\bigcup_{0\le i<n} A_i\Big| \le \sum_{0\le i<n} |A_i|.$$

Jos yhdisteen koko halutaan laskea tarkasti, käytetään inkluusio-ekskluusio -periaatetta eli summan ja erotuksen periaatetta:

$$\Big|\bigcup_{0\le i<n} A_i\Big| = \sum_{0\le i<n} |A_i| - \sum_{0\le i<j<n} |A_i\cap A_j| + \sum_{0\le i<j<k<n} |A_i\cap A_j\cap A_k| - \cdots. \tag{1}$$

Epäyhtälön oikean puolen voi nähdä tarkan summan (1) likiarvona: kun summa katkaistaan heti ensimmäisen termin jälkeen, saadaan yläraja. Yleisemmässä tapauksessa Bonferronin (tai Boolen–Bonferronin) epäyhtälö koskee kaikkia tällaisia likiarvoja: summa (1) plus- tai miinusmerkin kohdalta katkaistuna tuottaa vuorotellen ylä- ja alarajoja.

Bonferronin epäyhtälöstä on olemassa koko joukko erilaisia yleistyksiä erilaisiin käytännön tarpeisiin, ja esimerkiksi kirja [GS96] luettelee niitä paljon. Olen rajannut aihetta valitsemalla erään suurten datamäärien käsittelyyn liittyvän ongelman, jossa sopivasta yleistyksestä olisi apua. Esittelen tämän ongelman luvussa 3; lyhyesti sanottuna kyse on summan (1) arvioimisesta tilanteessa, jossa sitä ei voida katkaista siististi plus- tai miinusmerkin kohdalta, vaan jossa tunnetaan ne leikkaukset, joiden koko ylittää ennalta annetun rajan. Luku 3, joka siis liittyy tietojenkäsittelytieteeseen, on siinä määrin irrallinen kokonaisuus, että matemaatikkolukija voi hyvin ohittaa sen.

Luvussa 4 esittelen tuloksia, joita mieleni tekisi kutsua klassisiksi, vaikka osa niistä on ilmeisesti julkaistu vasta 1900-luvulla. Tässä yhteydessä pyrin esittämään vaihtelevia todistuksia siinä toivossa, että monipuolinen ajattelu ja työkalujen runsaus purevat ongelmaan.

Vanhoista tuloksista voi edetä kahteen suuntaan. Koska Bonferronin epäyhtälö on melko väljä, on luonnollista etsiä summan (1) termeille ± 1:n sijasta sellaisia kertoimia, että epäyhtälö tiukkenee. Toinen suunta, joka vaikuttaa kiinnostavammalta luvun 3 ongelman kannalta, on itse termien korjaaminen sellaisiksi, että niitä ei lasketa kaikkien indeksikombinaatioiden suhteen. Luku 5 käsittelee joitakin jälkimmäisen tyyppisiä tuloksia.

Luvussa 6 palataan luvun 3 ongelmaan. Luvun tulokset ovat siinä mielessä negatiivisia, että eräs suoraviivainen arvaus osoitetaan vääräksi ja ongelma jää erästä hyvin epätyydyttävää vastausta lukuunottamatta avoimeksi.

Hyvän arvauksen muotoileminen on osoittautunut hankalaksi, ja siinä uskoakseni piileekin ongelman vaikeus. Käytetty matematiikka on käsitteiltään varsin helppoa, pääosin jopa lukiotiedoin ymmärrettävää, joten uusienkaan tulosten todistaminen ei toivoakseni ole vaikeaa, kunhan sopiva konjektuuri saadaan muotoiltua.

Liitteessä A esitetään teknisiä lemmoja, joiden todistukset olisivat häiritsevän pitkiä muun tekstin keskellä. Todistukset on tuotettu tietokoneen avulla, joten ne eivät ole edes kovin kiinnostavia käsitteiden ymmärtämisen kannalta. Toisaalta on tietysti kiinnostavaa, että suuren osan tämän tutkielman lauseista voisi todistaa lähes täysin mekaanisesti.

Bonferronin epäyhtälö on luonteeltaan hyvin käytännöllinen. Summaa (1) varten on laskettava eksponentiaalisen monta lukua, joten summalle on pakko voida esittää arvioita. Tilastotieteilijät soveltavat Bonferronin epäyhtälöä yleensä hyvin yksinkertaisissa muodoissa, jotka sisältävät vain yksi tai kaksi ensimmäistä summaa. Syy on selvä: jos summista lasketaan mikään merkittävä osa, eksponentiaalinen räjähdys pääsee vaikuttamaan. Räjähdystä voi yrittää torjua luvun 5 tyyppisillä tuloksilla. Epäyhtälöä voi terävöittää muutamalla ylimääräisellä termillä, tai voi yrittää laatia sovelluksen kannalta sopivan erilaisen epäyhtälön.

Sen sijaan käsillä olevassa ongelmassa lasketaan joka tapauksessa joidenkin leikkausten kokoja. Tämä laskeminen on kallista, koska se vaatii tietokannan läpikäyntejä, joten lasketuista luvuista halutaan johtaa mahdollisimman paljon tietoa. Itse asiassa tarkan vastauksen laskeminen tietokannasta vaatisi vain yhden ylimääräisen läpikäynnin, mutta se halutaan välttää.

Voitaisiin kuvitella interaktiivinen järjestelmä, jolle esitetään kysymyksiä tietokannasta; jos käyttäjä saa heti summittaisen vastauksen virherajoineen, hän voi päättää käyttääkö resursseja tarkan vastauksen etsimiseen. Myös täysin automaattisessa järjestelmässä arvioista on hyötyä: tietokoneohjelma voi ensin käydä läpi jonkin suuren joukon potentiaalisesti kiinnostavia kysymyksiä ja arvioitujen vastausten perusteella valita kysymyksistä pienemmän joukon, jolle lasketaan tarkat vastaukset. Tiedon automaattiselle etsimiselle datasta on suurten tallennuskapasiteettien aikakaudella paljon kysyntää.

Koska etsin käytännössä sovellettavia ratkaisuja, asenteeni on jossain määrin konstruktivistinen. Esimerkiksi lauseen 6 (sivulla 15) voi todistaa lyhyemminkin: muunnetaan propositiokaava φ disjunktiiviseen normaalimuotoon, ja tehdään muunnos niin pitkälle, että kaikki disjunktit ovat toisensa poissulkevia; jokaiseen disjunktiin voidaan soveltaa inkluusio-ekskluusio -kaavaa ja laskea lopputulokset yhteen. Tämän normaalimuodon käytön takia todistus ei ole käytännöllinen, millä tarkoitan ettei siitä voi kirjoittaa järkevää tietokoneohjelmaa. Kaavan toteutuvuus on nimittäin tunnetusti NP-täydellinen ongelma, mutta disjunktiivisessa normaalimuodossa olevan kaavan toteutuvuus on hyvin helppo tarkistaa. Siis tähän normaalimuotoon muuntamisen on oltava vaikea tehtävä.

Toinen esimerkki on Rényin lause (lause 13 sivulla 33), joka sallisi useimpien tulosten todistamisen jonkin verran vähemmällä vaivalla. Ennen kaikkea tulokset eivät olisi enää vain kombinatorisia, vaan niiden pätevyys ulottuisi paljon laajemmalle, kaikkiin todennäköisyysavaruuksiin. Sovelluksen kannalta tällä yleisyydellä ei kuitenkaan ole käyttöä. Mielestäni lauseet ovat helpompia hahmottaa ja herättävät vähemmän epäilyksiä soveltajassa, kun ne koskevat äärellisiä joukkoja. Todistukset ovat toki siinä mielessä turhan pitkiä, että Rényin lause poistaisi kustakin yhden askeleen, jossa vaihdetaan summausjärjestystä. Vastapainona todistukset ovat uskoakseni helpompia lukea. Probabilistisesti suuntautunut lukija voi joka tapauksessa helposti korottaa tulosten yleisyystasoa.

2 Käsitteitä ja merkintöjä

Seuraavassa luetellaan tutkielmassa käytettyjä merkintöjä. Ajatus on, että lukija tuntee varsinaiset asiat ja voi tarkistaa tästä, millä tavalla niitä käytetään tässä työssä. Seuraavan luettelon perusteella on tuskin mahdollista oppia käsitteitä alusta lähtien. Suuri osa asioista löytyy esimerkiksi kirjasta [GKP90]. Lisäksi luvussa 5 käytetään verkkoteoriaa, jonka käsitteistä on vastaava luettelo kappaleessa 5.1.

Kokonaislukujen joukkoa merkitään symbolilla $\mathbb{Z}$, positiivisten kokonaislukujen joukkoa symbolilla $\mathbb{Z}^+$, reaalilukujen joukkoa symbolilla $\mathbb{R}$ ja positiivisten reaalilukujen joukkoa symbolilla $\mathbb{R}^+$. Luonnollisia lukuja ovat kaikki ei-negatiiviset kokonaisluvut eli $\mathbb{N} = \mathbb{Z}^+ \cup \{0\}$.

Kun $n \in \mathbb{Z}$, merkitään $[n] = \{\, j \in \mathbb{N} \mid j < n \,\}$. Esimerkiksi $[-1] = [0] = \emptyset$ ja $[2] = \{0, 1\}$.

Äärellisen joukon X *koolle* (kardinaliteetille, alkioiden lukumäärälle) käytetään enimmäkseen merkintää $|X|$, mutta silloin kun pystyviiva olisi vaarassa sekaantua muiden merkintöjen kanssa, merkitään $\#X$. Joukon X kaikkien osajoukkojen perhe, *potenssijoukko*, on $\mathcal{P}(X)$.

Kun $n \in \mathbb{Z}$, joukon X niitä osajoukkoja, joissa on tasan n alkiota, sanotaan X:n *n-osajoukoiksi*. Joukon X n-osajoukkojen perhettä sanotaan X:n *n-tasojoukoksi* ja merkitään $X^{(n)}$. Toisin sanoen

$$X^{(n)} = \{\, Y \subseteq X \mid \#Y = n \,\}.$$

Binomikertoimet määritellään kaikille $n \in \mathbb{R}$, $p \in \mathbb{Z}$ kaavalla

$$\binom{n}{p} = \begin{cases} \dfrac{n(n-1)\cdots(n-p+1)}{p(p-1)\cdots 1}, & p \geq 0, \\ 0, & p < 0. \end{cases}$$

Jos $n, p \in \mathbb{N}$, saadaan yksinkertaisempi muotoilu

$$\binom{n}{p} = \frac{n!}{(n-p)!p!},$$

missä huutomerkki tarkoittaa *kertomaa:* $x! = x(x-1)\cdots 1$, $0! = 1$.

Seuraavat binomikertoimia koskevat lauseet oletetaan tunnetuiksi.

Lause 1. *Olkoot $n \in \mathbb{R}$, $p, q \in \mathbb{Z}$. Silloin*

$$
\begin{array}{lll}
\textbf{(a)} & \binom{n}{0} = 1, & \\
\textbf{(b)} & \binom{n}{p} = 0, & 0 \le n < p,\ n \in \mathbb{N}, \\
\textbf{(c)} & \binom{n}{p} = \binom{n}{n-p}, & n \in \mathbb{N}, \\
\textbf{(d)} & \binom{n}{p} = \binom{n-1}{p-1} + \binom{n-1}{p}, & \\
\textbf{(e)} & \sum_{j \in \mathbb{Z}} \binom{n}{j} = 2^n, & n \in \mathbb{N}, \\
\textbf{(f)} & \binom{n}{p}\binom{p}{q} = \binom{n}{q}\binom{n-q}{p-q}, & \\
\textbf{(g)} & \binom{n}{p} = \frac{n}{p}\binom{n-1}{p-1}, & p \in \mathbb{Z} \setminus \{0\}.
\end{array}
$$

Lause 2. *Olkoot $n, p \in \mathbb{N}$. Silloin*

$$\binom{n}{p} = \left|[n]^{(p)}\right|$$

eli jokaisella n-alkioisella joukolla on $\binom{n}{p}$ eri p-osajoukkoa.

Lause 3 (Binomilause). *Kaikilla $x, y \in \mathbb{R} \setminus \{0\}$, $x \neq -y$, ja $n \in \mathbb{N}$, pätee*

$$(x+y)^n = \sum_{p=0}^{n} \binom{n}{p} x^p\, y^{n-p}.$$

Koska tässä työssä käsitellään enimmäkseen luonnollisia lukuja, sovitaan että $0^0 = 1$, jolloin binomilause pätee kaikilla $x, y \in \mathbb{R}$.

Bonferroni-tyyppisiä epäyhtälöitä käsiteltäessä voitaisiin usein puhua *monijoukoista*, epämuodollisesti sanottuna "joukoista, joihin alkio voi kuulua useita kertoja". Tarkastellaan esimerkiksi seuraavaa Bonferronin epäyhtälön tunnettua erikoistapausta: Olkoon $\mathcal{A}$ perhe joukkoja. Joukkojen yhdisteen koko on enintään joukkojen yksittäisten kokojen summa,

$$\left|\bigcup \mathcal{A}\right| \le \sum_{X \in \mathcal{A}} |X|\,. \tag{2}$$

Jos jokin joukoista $X \in \mathcal{A}$ lasketaan mukaan useita kertoja, epäyhtälön vasen puoli pysyy samana ja oikea voi vain kasvaa. Perheen $\mathcal{A}$ voisi siis sallia olevan monijoukko.

Monijoukko voitaisiin määritellä esimerkiksi funktiona jostain perusjoukosta luonnollisille luvuille. Merkintöjen ja käsitteiden yksinkertaistamiseksi luovutaan kuitenkin monijoukoista ja puhutaan niiden sijasta *jonoista*. Esimerkiksi epäyhtälössä (2) voidaan ajatella $\mathcal{A}$:n olevan jono joukkoja, kunhan merkinnät $\bigcup \mathcal{A}$ ja $X \in \mathcal{A}$ yleistetään luonnollisella tavalla jonoille.

Muodollisesti jonolla tarkoitetaan paria $\langle I, f \rangle$, missä I on jokin lineaarisesti järjestetty indeksijoukko ja f joukossa I määritelty kuvaus; käytännössä I:ksi on kätevää valita jokin joukko luonnollisia lukuja. Jos $X = \langle I, f \rangle$ on jono, jonoa $Y = \langle I', f' \rangle$ sanotaan X:n osajonoksi, jos $I' \subseteq I$ ja $f' = f \restriction I'$, ts. f' on f:n rajoittuma joukkoon I'. Osajonon Y voi siis ajatella sisältävän joitakin X:n alkioita, jotka eivät välttämättä ole peräkkäisiä mutta ovat samassa järjestyksessä kuin X:ssä. Kirjoitetaan hieman merkintöjä väärinkäyttäen $Y \subseteq X$, ja yleistetään samoin merkinnät $x \in X$, $\mathcal{P}(X)$ ja $X^{(n)}$ luonnollisella tavalla jonoille X.

On huomattava, että jos jonot kirjoitetaan normaaliin tapaan luettelemalla alkiot järjestyksessä, toistetut alkiot voivat tuottaa sekaannuksia. Esimerkiksi jonolla (a, a, b) on kolme kaksialkioista osajonoa, joista kahdella on sama esitys (a, b) tätä merkintää käytettäessä. Ongelma johtuu tietysti siitä, että indeksijoukko jätetään kirjoittamatta.

Eräs tapa tuottaa jonoja on soveltaa jotain funktiota toisten jonojen alkioihin. Jos $X = \langle I, f \rangle$ on jono ja g on sen alkioille määritelty funktio, merkitään

$$(\, g(x) \mid x \in X \,) = \langle I, g \circ f \rangle.$$

Merkintä tarkoittaa siis jonoa, jonka alkiot saadaan funktiolla g X:n alkioista. Edelleen, jos P on predikaatti, X:n P:llä valittu osajono

$$(\, x \in X \mid P(x) \,) = \langle I', f \restriction I' \rangle$$

on jono, jolle $I' = \{ \, i \in I \mid P(f(i)) \, \}$.

Koska jonoja käytetään tässä lähinnä yksinkertaistamaan esitystä sillä, että vältetään monijoukoista puhuminen, niihin yleensä sisältyvä ajatus *järjestyksestä* on vähemmän tärkeä. Jonoja määriteltäessä jätetään siksi usein huomiotta kysymys jonon yksikäsitteisyydestä järjestyksen suhteen, ja tyydytään ajattelemaan, että indeksijoukolla on jokin kanoninen järjestys.

Yhteen- ja kertolaskusymboleita $\sum$ ja $\prod$ käytetään tavanomaisesti kahdella tavalla. Ensimmäisestä tavasta on esimerkkinä epäyhtälön (2) oikea puoli: symbolin alle kirjoitetaan inkluusiorelaatio ja oikealle puolelle lauseke, jossa esiintyvä vapaa muuttuja ajatellaan relaation sitomaksi. Tämän merkinnän erikoistapauksena peräkkäisten kokonaislukujen joukosta ilmoitetaan usein vain pienin ja suurin alkio, esimerkiksi

$$\sum_{j=0}^{n} j^2 = 0^2 + 1^2 + \cdots + n^2.$$

Toinen tapa on ajatella symbolit operaattoreiksi, jotka operoivat joukkoon tai jonoon:

$$n! = \prod \left\{ j \;\middle|\; j \in \mathbb{Z}^+, \, j \le n \right\}, \quad |X| = \sum \left(1 \;\middle|\; x \in X \right).$$

Edellä toinen esimerkki osoittaa myös, miksi ei riitä määritellä operaattoreita $\sum$ ja $\prod$ vain joukoille: joukossa $\{\, 1 \mid x \in X \,\}$ on tietysti vain yksi alkio, mutta on mielekästä laskea yhteen samaakin alkiota useita kertoja. Näin yksinkertaisissa tapauksissa ensimmäisen merkinnän mainittu erikoistapaus olisi toki kätevämpi, mutta varsinkin joissain teknisissä todistuksissa esiintyy summia, joissa vapaan muuttujan sitova ehto on monimutkaisempi.

Edelliset merkintäsopimukset on hyödyllistä yleistää myös yhdisteille ja leikkauksille. Epäyhtälön (2) vasemmalla puolella käytetään yhdisteoperaattoria $\bigcup$ joukkoon $\mathcal{A}$. Sovitaan edelleen, että $\bigcup$ ja $\bigcap$ voivat operoida myös jonoihin.

3 Tietämyksen muodostaminen assosiaatiosäännöillä

Tietämyksen muodostaminen (data mining, knowledge discovery) on tutkimusalue, joka yhdistää koneoppimisen, tilastotieteen ja tietokantatutkimuksen menetelmiä. Tavoite on lyhyesti sanottuna löytää suurista datamääristä kiinnostavaa uutta tietoa.

Eräs tietämyksen muodostamisen menetelmä on *assosiaatiosääntöjen* etsiminen. Esimerkiksi ruokakaupan ostostietokannasta voidaan löytää automaattisesti sellaisia sääntöjä kuin

$$\{\text{olut}, \text{sinappi}\} \Rightarrow \{\text{makkara}\}.$$

Tämän säännön tulkinta on "ostoskoreissa, joissa on olutta ja sinappia, on usein myös makkaraa".

Tässä luvussa määritellään assosiaatiosäännöt ja joitakin taustakäsitteitä. Sen jälkeen esitetään algoritmi, joka löytää tietokannasta kaikki eräässä mielessä kiinnostavat assosiaatiosäännöt. Lopuksi yritetään yleistää assosiaatiosääntöjä ja esitetään, mitä hyötyä eräänlaisesta Bonferronin epäyhtälön muunnelmasta olisi yleistettyjen sääntöjen etsimisessä.

Assosiaatiosäännöt määriteltiin julkaisussa [AIS93]. Artikkelissa [AMS$^+$96] määritelmää laajennettiin hieman ja esitettiin nopea algoritmi sääntöjen etsimiseen. Yleisempien sääntöjen ajatus on peräisin konferenssijulkaisusta [MT96].

3.1 Relaatiot

Tietämyksen muodostamisessa oletetaan, että käsiteltävänä on paljon dataa. Tietojenkäsittelytieteessä suuria eräällä tavalla organisoituja datamääriä kutsutaan relaatiotietokannoiksi.

Tämän työn tarkoituksia varten riittävät seuraavat tavallista rajoittuneemmat määritelmät. Käytännön tietokannoissa on useita relaatioita, jotka sisältävät mm. numeroita, merkkijonoja ja viittauksia toisiin relaatioihin. Tällainen laajempi määritelmä esitetään lähes kaikissa tietokantoja käsittelevissä oppikirjoissa, esimerkiksi [EN94]:n luvussa 6.

Määritelmä 1. *Relaatioskeema* S on joukko $\{A_1, A_2, \ldots, A_k\}$. Sen alkioita sanotaan *attribuuteiksi*. Skeeman *aste* $\deg S$ on attribuuttien lukumäärä k.

Määritelmä 2. Relaatioskeeman S toteuttava *relaatio* R on äärellinen jono, jonka alkiot, *monikot*, ovat S:n osajoukkoja. Relaation *aste* $\deg R$ on sama kuin sen skeeman aste $\deg S$, ja relaation *koko* on sen monikkojen lukumäärä $|R|$.

Relaation monikot voidaan samaistaa karakterististen funktioidensa kanssa: ajatellaan, että monikko liittää jokaiseen attribuuttiin arvon 0 tai 1. Monikkoa voidaan pitää luonnollisella tavalla myös k-alkioisena jonona. Jos esimerkiksi $k = 3$, relaatio sisältää kolmikkoja; "monikko" (engl. tuple) on tämäntapaisten sanojen yleistys.

Taulukko 1 esittää karakteristisen funktion avulla erästä relaatiota, jonka aste on 5 ja koko 6. Esimerkiksi ruokakaupan yhden vuoden ostoskorirelaation aste on luultavasti satoja tai tuhansia ja koko kymmenistä tuhansista ylöspäin, mutta kussakin monikossa on

Olut	Makkara	Sinappi	Tomaatti	Keräkaali
0	1	1	0	0
0	1	1	0	0
1	1	0	0	0
0	0	0	1	1
1	0	0	1	0
1	1	0	1	0

Taulukko 1: Esimerkki relaatiosta

melko vähän alkioita. Eri sovelluksissa aste vaihtelee kymmenistä tuhansiin, ja koko voi olla miljoonia tai jopa satoja miljoonia.

Vaikka aste olisi hyvin suuri, monikot voivat olla pieniä. Jos relaatiolla on tällainen harva rakenne, sitä ei tietenkään kannata tallentaa sellaisenaan nollina ja ykkösinä. Relaatiota ei välttämättä ole edes varsinaisesti olemassa, vaan kyselyihin vastataan laskemalla jostain muusta tiedosta.

Logiikassa skeeman S toteuttavaa relaatiota kutsuttaisiin äärelliseksi malliksi, jonka aakkosto S sisältää vain yksipaikkaisia relaatiosymboleita.

3.2 Kattavat hahmot

Assosiaatiosäännöt perustuvat *hahmoihin*[1]: ajatus on, että hahmo sisältää piirteitä, jotka sopivat joihinkin monikoihin. Seuraavassa hahmo määritellään yksinkertaisesti attribuuttijoukoksi, jolloin se sopii niihin monikoihin, jotka sisältävät kaikki hahmon attribuutit. Yleisemmistä tietokannoista puhuttaessa hahmot voitaisiin määritellä osajoukkojen sijasta jonkinlaisiksi säännöiksi. Toisaalta käsillä olevan sovelluksen tavoitteena on löytää koneellisesti relaatioon sopivia hahmoja (ja niiden avulla assosiaatiosääntöjä). Potentiaalisten hahmojen lukumäärä voisi kasvaa kovin suureksi, jos sallittaisiin sellaiset hahmot kuin "attribuutin Nimi arvo sisältää J-kirjaimen ja attribuutin Pituus arvo on suurempi kuin 179,4". Siksi voi olla tarkoituksenmukaista valita ennalta joitakin kiinnostavia hahmotyyppejä ja lisätä relaatioon niitä vastaavia $\{0,1\}$-arvoisia attribuutteja.

Määritelmä 3. Olkoon S relaatioskeema ja R sen toteuttava relaatio. Skeeman S osajoukkoja sanotaan S:n (tai R:n) *hahmoiksi*. Hahmo $T \subseteq S$ *sopii* niihin relaation R monikoihin v, joille $T \subseteq v$. Hahmon *kattavuus* relaatiossa R on niiden R:n monikkojen osuus, joihin hahmo sopii,

$$\mathrm{freq}_R(T) = \frac{|(\, v \in R \mid v \supseteq T\,)|}{|R|}.$$

[1]Englanninkielisessä kirjallisuudessa puhutaan usein tiheistä joukoista (frequent sets). Mielestäni sana "hahmo" kuvaa käsitettä hyvin ja aiheuttaa vähemmän sekaannuksia, kun muutenkin on puhe joukoista.

Merkinnästä jätetään pois relaatiota osoittava alaindeksi, jos relaatio on selvä asiayhteydestä. Jos on annettu jokin kattavuusraja ρ, sen ylittäviä hahmoja sanotaan *ρ-kattaviksi* tai lyhyesti *kattaviksi*.

Esimerkiksi taulukon 1 relaatiossa hahmon $\{\text{Olut}, \text{Keräkaali}\}$ kattavuus on 0, ja $\{\text{Makkara}\}$ on kattavuudeltaan suurin epätyhjä hahmo: $\text{freq}(\{\text{Makkara}\}) = 2/3$. Tyhjälle hahmolle pätee triviaalisti $\text{freq}(\emptyset) = 1$.

On selvää, että jos hahmoon lisätään attribuutteja, sen kattavuus ei voi kasvaa. Tätä monotonisuusominaisuutta käytetään kattavien hahmojen etsinnässä, joten kirjoitetaan se lauseeksi:

Lause 4. *Olkoon R relaatio. Jos T ja T' ovat R:n hahmoja, joille pätee $T \subseteq T'$, hahmojen kattavuuksille on voimassa*

$$\text{freq}_R(T') \leq \text{freq}_R(T).$$

Todistus. Olkoot $R_T = (\, v \in R \mid v \supseteq T \,)$ ja $R_{T'} = (\, v \in R \mid v \supseteq T' \,)$. Kaikilla $v \in R_{T'}$ pätee $T \subseteq T' \subseteq v$ ja siis $v \in R_T$. Toisin sanoen $R_{T'} \subseteq R_T$, mistä väite seuraa. □

3.3 Assosiaatiosäännöt

Määritelmä 4. Olkoon R relaatio. Jos $X, Y \subseteq S$ ovat R:n erillisiä epätyhjiä hahmoja,

$$X \Rightarrow Y$$

on (eräs) R:n *assosiaatiosääntö*. Tämän säännön *tuki* relaatiossa R on

$$\text{supp}_R(X \Rightarrow Y) = \text{freq}_R(X \cup Y)$$

ja *luotettavuus* relaatiossa R on

$$\text{conf}_R(X \Rightarrow Y) = \frac{\text{freq}_R(X \cup Y)}{\text{freq}_R(X)}.$$

Jos $\text{freq}_R(X) = 0$, määritellään luotettavuudeksi mielivaltaisesti 0.

Huomautus 1. Luotettavuuden kaava voi näyttää kummalliselta: osoittajassa esiintyvä joukko (hahmo) $X \cup Y$ on nimittäjän joukkoa X suurempi. Tässä ei kuitenkaan lasketa joukkojen X ja $X \cup Y$ kokoja vaan niitä vastaavien relaation osajonojen kokoja, jolloin lauseesta 4 seuraa, että luotettavuus on aina välillä $[0, 1]$.

Säännön $X \Rightarrow Y$ intuitiivinen merkitys on "jos hahmo X sopii johonkin monikkoon, myös hahmo Y sopii siihen." Tuki ilmaisee sen, kuinka suureen osaan relaatiota sääntö perustuu, luotettavuus sen, kuinka tarkasti sääntö pitää paikkansa. Tavallinen tehtävä tietämyksen muodostamisessa on löytää relaatiosta kaikki riittävästi tuetut riittävän luotettavat assosiaatiosäännöt.

Nro	Sääntö	Tuki	Luotettavuus
1	$\{\mathrm{Sinappi}\} \Rightarrow \{\mathrm{Makkara}\}$	$1/3$	1
2	$\{\mathrm{Keräkaali}\} \Rightarrow \{\mathrm{Tomaatti}\}$	$1/6$	1
3	$\{\mathrm{Olut}, \mathrm{Tomaatti}\} \Rightarrow \{\mathrm{Makkara}\}$	$1/6$	$1/2$
4	$\{\mathrm{Makkara}, \mathrm{Sinappi}\} \Rightarrow \{\mathrm{Olut}\}$	0	0

Taulukko 2: Esimerkkejä assosiaatiosäännöistä

Esimerkki 1. Taulukossa 2 on lueteltu muutamia taulukon 1 relaation assosiaatiosääntöjä. Sääntö 1 on hyödyllisempi kuin sääntö 2, vaikka kummankin luotettavuus on sama, koska sääntö 1 kertoo suuremmasta relaation osasta. Sääntö 2 on hyödyllisempi kuin sääntö 3, koska sääntö 3 on vähemmän luotettava. Koska säännön 4 tuki on 0, se ei kerro relaatiosta muuta kuin että attribuutit Olut, Sinappi ja Makkara eivät esiinny samoissa monikoissa.

Muutkin kuin tuetut ja luotettavat säännöt voisivat olla kiinnostavia: jos kaupan ostoskoritietokannassa esimerkiksi säännön 4 tuki ja luotettavuus ovat pienet, kauppias voisi päätellä, että makkaraa sinapin kanssa syövät eivät ole oluen ystäviä. Assosiaatiosääntöjen etsimisalgoritmit eivät kuitenkaan löydä tällaista tietoa; tarve poistaa tämä rajoitus johtaa kappaleessa 3.6 käsiteltäviin yleisempiin sääntöihin.

3.4 Kattavien hahmojen löytäminen

Tuettujen ja luotettavien sääntöjen etsimisalgoritmi perustuu kattaviin hahmoihin. Mahdollisia hahmoja on tietysti eksponentiaalisen paljon suhteessa attribuuttien määrään, joten kaikkien hahmojen tarkasteleminen on käytännössä toivoton tehtävä. Tässä auttaa tietysti kattavuuden monotonisuus: asetetaan ennalta jokin kattavuusraja ja hylätään mahdollisimman paljon sen alittavia hahmoja lauseen 4 avulla. Aluksi kannattaa laskea yksialkioisten hahmojen kattavuudet, sillä jos jokin niistä jää rajan alle, sen avulla voidaan sulkea pois useita suurempia hahmoja. Edetään seuraavaksi niihin kaksialkioisiin hahmoihin, joita ei ole suljettu pois, ja niin edelleen. Ajatus on siis etsiä kattavat hahmot tasoittain, joten määritellään kattavien hahmojen taso.

Määritelmä 5. Olkoon R skeeman S toteuttava relaatio, $\rho \in [0,1]$ kattavuusraja ja $\ell \in \mathbb{Z}^+$. Silloin $\mathcal{F}_\ell(R,\rho)$ on niiden R:n hahmojen joukko, joissa on ℓ alkiota ja joiden kattavuus on vähintään ρ,

$$\mathcal{F}_\ell(R,\rho) = \{\, T \in S^{(\ell)} \mid \mathrm{freq}_R(T) \geq \rho \,\}.$$

Tässä $S^{(\ell)}$ tarkoittaa S:n ℓ-tasoa: ehdon $T \in S^{(\ell)}$ voisi kirjoittaa myös muotoon $T \subseteq S$, $|T| = \ell$.

Kun tason ℓ kattavat hahmot tunnetaan, tiedetään että tason $\ell + 1$ jokaisen kattavan hahmon ℓ-osajoukot ovat välttämättä tunnettujen hahmojen joukossa. Näin saadaan menetelmä, jolla perheen $\mathcal{F}_\ell(R, \rho)$ hahmoista voidaan laskea ehdokkaat perheen $\mathcal{F}_{\ell+1}(R, \rho)$ hahmoiksi.

Määritelmä 6. Olkoon U joukko ja $\ell \in \mathbb{Z}^+$. Määritellään funktio $\mathcal{C}_{\ell+1} : \mathcal{P}(U^{(\ell)}) \to \mathcal{P}(U^{(\ell+1)})$ kaavalla

$$\mathcal{C}_{\ell+1}(\mathcal{X}) = \{\, Y \in U^{(\ell+1)} \mid Y^{(\ell)} \subseteq \mathcal{X} \,\}.$$

Toisin sanoen $\mathcal{C}_{\ell+1}$ kuvaa perheen $\mathcal{X}$, jonka alkiot ovat ℓ-alkioisia U:n osajoukkoja, perheeksi $\mathcal{C}_{\ell+1}(\mathcal{X})$, joka sisältää kaikki ne U:n osajoukot Y, joille pätee: ensinnäkin $|Y| = \ell+1$, ja toiseksi jokainen Y:n ℓ-alkioinen osajoukko kuuluu perheeseen $\mathcal{X}$.

Lause 5. *Olkoon R skeeman S toteuttava relaatio, $\rho \in [0, 1]$ kattavuusraja ja $\ell \in \mathbb{Z}^+$. Silloin tason $\ell+1$ kattavat hahmot kuuluvat siihen perheeseen, joka saadaan funktiolla $\mathcal{C}_{\ell+1}$ tason ℓ kattavista hahmoista, eli*

$$\mathcal{F}_{\ell+1}(R, \rho) \subseteq \mathcal{C}_{\ell+1}(\mathcal{F}_\ell(R, \rho)),$$

missä funktion $\mathcal{C}_{\ell+1}$ määritelmässä mainittu perusjoukko U on S.

Todistus. Olkoon $T \in \mathcal{F}_{\ell+1}(R, \rho)$ kattava hahmo. Silloin lauseen 4 perusteella jokainen T:n osajoukko on myös kattava hahmo. Erityisesti kaikilla $T' \in T^{(\ell)}$ pätee $\mathrm{freq}_R(T') \geq \mathrm{freq}_R(T) \geq \rho$, ja siis $T^{(\ell)} \subseteq \mathcal{F}_\ell(R, \rho)$. Väite seuraa määritelmän 6 nojalla. $\square$

Tämän lauseen avulla voidaan nyt toteuttaa algoritmi kaikkien kattavien hahmojen etsimiseksi. Algoritmi A ottaa syötteenä skeeman S, relaation R ja kattavuusrajan $\rho \in [0, 1]$ ja tulostaa kaikki R:n epätyhjät hahmot, joiden kattavuus on vähintään ρ. Se käyttää apumuuttujina joukkoja $\mathcal{E}_\ell$, jotka sisältävät kunkin tason ℓ ehdokkaat kattaviksi hahmoiksi, ja laskee kattavat hahmot muuttujiin $\mathcal{F}_\ell$.

Myöhempää varten kannattaa huomata, että eräänlaisena sivutuotteena algoritmi A tulee laskeneeksi myös sellaisten hahmojen tiheydet, jotka eivät ole kattavia mutta joiden kaikki osajoukot ovat.

Algoritmi A

$\mathcal{E}_1 := S^{(1)}$
$\ell := 1$
toista kunnes $\mathcal{E}_\ell = \emptyset$:
    $\mathcal{F}_\ell := \{\, T \in \mathcal{E}_\ell \mid \mathrm{freq}_R(h) \geq \rho \,\}$
    $\mathcal{E}_{\ell+1} := \mathcal{C}_{\ell+1}(\mathcal{F}_\ell)$
    $\ell := \ell + 1$
toista kaikille $\ell \geq 1$:
    toista kaikille $T \in \mathcal{F}_\ell$:
        tulosta hahmo T ja kattavuus $\mathrm{freq}_R(T)$

Rivillä 1 valitaan ensimmäisen tason ehdokkaiksi kaikki yksialkioiset hahmot, ja sitten ryhdytään iteroimaan ehdokasmenettelyä. Jokaisella iteraatiolla valitaan ehdokkaista todelliset kattavat hahmot rivillä 4 ja lasketaan ehdokkaat seuraavalle iteraatiolle rivillä 5. Rivin 4 laskentaan riittää tavallisesti yksi tietokannan läpikäynti: tarkistetaan vuorotellen kustakin monikosta, mitkä hahmot sopivat siihen, ja lisätään näiden hahmojen kattavuuslaskureita.

Rivillä 5 käytettävän funktion $\mathcal{C}_{\ell+1}$ järkevä laskenta ei ole itsestäänselvää. Suoraviivainen ja tehoton tapa on käydä läpi kaikki joukon $S^{(\ell+1)}$ hahmot ja tarkistaa määritelmän 6 ehto. Parempi idea on pitää joukon $\mathcal{F}_\ell$ hahmot muistissa leksikografisessa järjestyksessä ja käyttää sitä hyväksi. Seuraava algoritmi laskee funktion $\mathcal{C}_{\ell+1}$ tämän idean avulla.

Algoritmi $\mathbf{A_1}$

toista kaikille $X \in \mathcal{F}_\ell$:
    toista kaikille $Y \in \mathcal{F}_\ell$, joille
    $Y > X$ ja X:n ja Y:n $\ell - 1$ ensimmäistä alkiota ovat samat:
        toista kaikille $Z \in (X \cup Y)^{(\ell)}$:
            jos Z ei ole joukossa $\mathcal{F}_\ell$, siirry seuraavaan Y
        tulosta $X \cup Y$

Algoritmi A_1 vaikuttaa selvästi nopeammalta kuin edellä kuvattu naiivi menetelmä. Tarkempi analyysi sivuutetaan tässä yhteydessä.

3.5 Assosiaatiosääntöjen löytäminen

Assosiaatiosääntöjen etsimisalgoritmi on suhteellisen helppo, kun kattavat hahmot tunnetaan. Algoritmi B ottaa syötteenä skeeman S, relaation R, minimituen σ ja minimiluottamuksen τ. Se tulostaa kaikki assosiaatiosäännöt, joiden tuki on vähintään σ ja luottamus vähintään τ.

Algoritmi B

1 laske σ-kattavien hahmojen joukko $\mathcal{F}$ algoritmilla A
2 toista kaikille $X \in \mathcal{F}$:
3   toista kaikille $Y \subsetneq X$, $Y \neq \emptyset$:
4   jos $\mathrm{freq}_R(X) / \mathrm{freq}_R(X \setminus Y) \geq \tau$, niin:
5   tulosta sääntö $X \setminus Y \Rightarrow Y$, tuki $\mathrm{freq}_R(X)$ ja luottamus $\mathrm{freq}_R(X) / \mathrm{freq}_R(X \setminus Y)$

Rivillä 1 käytetään algoritmia A, jolle annetaan syötteenä S, R ja σ. Jos assosiaatiosäännön tuki on vähintään σ, siinä esiintyvien attribuuttien muodostaman hahmon kattavuus on tuen määritelmän perusteella vähintään σ. Siis rivillä 4 tulevat testatuiksi kaikki riittävän tuetut assosiaatiosäännöt, ja niistä riittävän luotettavat tulostetaan rivillä 5. Rivin 1 jälkeen algoritmin B ei tarvitse käyttää tietokantaa lainkaan, sillä riveillä 4 ja 5 tarvitaan lauseen 4 perusteella vain joukon $\mathcal{F}$ hahmojen kattavuuksia.

3.6 Propositiosäännöt

Kappaleen 3.3 lopussa tarkasteltiin lyhyesti erästä assosiaatiosääntöihin liittyvää rajoitusta: jos tietokannasta etsitään vain kattavia ja luotettavia sääntöjä, ei saada tietoa attribuuttien monimutkaisemmista suhteista. Ongelman voi ajatella johtuvan siitä, että säännöissä on implikaationuolen kummallakin puolella konjunktio. Jos sallittaisiin monimutkaisemmat lausekkeet, voitaisiin puhua esimerkiksi taulukon 1 tietokannan säännöstä

$$\text{Olut} \Rightarrow \neg\text{Sinappi},$$

jonka tuki olisi 1/2 ja luotettavuus 1. Seuraavassa määritellään laajempi luokka sääntöjä ja tutkitaan niiden luotettavuuksien arvioimista.

Siinä missä assosiaatiosäännöt perustuvat hahmoihin, yleisemmät säännöt rakennetaan kaavoista. Propositiologiikan atomikaavoja vastaavat tässä S:n attribuutit, ja yleisemmät kaavat rakennetaan tavalliseen tapaan rekursiivisesti.

Määritelmä 7. Relaatioskeeman S *kaavoja* ovat

1. attribuutit $A \in S$,
2. muotoa $\neg\varphi$ olevat merkkijonot, missä φ on S:n kaava,
3. muotoa $(\varphi \wedge \psi)$ olevat merkkijonot, missä φ ja ψ ovat S:n kaavoja ja
4. muotoa $(\varphi \vee \psi)$ olevat merkkijonot, missä φ ja ψ ovat S:n kaavoja.

Hahmojen sopivuutta vastaa kaavojen totuus, joka määritellään myös samaan tapaan kuin propositiologiikassa.

Määritelmä 8. Olkoon R skeeman S toteuttava relaatio ja $v \in R$ monikko. S:n kaavan *totuus* monikolle v määritellään seuraavasti.

1. Jos $A \in R$, kaava A on tosi monikolle v, jos ja vain jos $A \in v$.
2. Kaava $\neg\varphi$ on tosi monikolle v, jos ja vain jos kaava φ ei ole tosi monikolle v.
3. Kaava $(\varphi \wedge \psi)$ on tosi monikolle v, jos ja vain jos kumpikin kaavoista φ ja ψ on tosi monikolle v.
4. Kaava $(\varphi \vee \psi)$ on tosi monikolle v, jos ja vain jos ainakin toinen kaavoista φ ja ψ on tosi monikolle v.

Jos kaava φ on tosi monikolle v, merkitään $v \vDash \varphi$. Muuten sanotaan, että φ on epätosi v:lle, $v \nvDash \varphi$. Skeeman S kaavojen φ ja ψ sanotaan olevan (loogisesti) *ekvivalentteja*, jos kaikilla $v \subseteq S$ pätee $v \vDash \varphi$ täsmälleen silloin, kun $v \vDash \psi$.

Koska tämä totuusmääritelmä vastaa logiikassa tavanomaisesti käytettyä, seuraavassa sovelletaan propositiologiikan tunnettuja lauseita kaavoihin. Niistä seuraa, että kaavojen kirjoitusasua voidaan yksinkertaistaa jättämällä tarpeettomat sulkeet kirjoittamatta. Sovitaan myös, että negaatiosymboli $\neg$ liittyy aina lyhimpään mahdolliseen seuraavaan alikaavaan, mutta ei kiinnitetä merkeille $\vee$ ja $\wedge$ keskinäistä prioriteettia. Symbolia $\rightarrow$ voitaisiin käyttää tavanomaiseen tapaan lyhennysmerkintänä, niin että kaava $\varphi \rightarrow \psi$ merkitsee kaavaa $\neg\varphi \vee \psi$, mutta tälle ei ole seuraavassa tarvetta. On syytä huomata, että kaksinkertaista nuolta $\Rightarrow$ käytetään sääntöjen, ei kaavojen, tasolla.

Määritelmä 9. Kaavan φ *kattavuus* relaatiossa R on niiden monikoiden osuus, joille φ on tosi,

$$\mathrm{freq}_R(\varphi) = \frac{|(\, v \in R \mid v \models \varphi \,)|}{|R|}.$$

Kaavoista rakennetaan nyt propositiosäännöt samalla tavalla kuin hahmoista assosiaatiosäännöt.

Määritelmä 10. Olkoot φ ja ψ relaatioskeeman S kaavoja. Silloin $\varphi \Rightarrow \psi$ on S:n *propositiosääntö*. Tämän säännön *tuki* S:n toteuttavassa relaatiossa R on

$$\mathrm{supp}_R(\varphi \Rightarrow \psi) = \mathrm{freq}_R(\varphi \wedge \psi),$$

ja *luotettavuus* relaatiossa R on

$$\mathrm{conf}_R(\varphi \Rightarrow \psi) = \frac{\mathrm{freq}_R(\varphi \wedge \psi)}{\mathrm{freq}_R(\varphi)}. \tag{3}$$

Propositiosäännöt ovat luonnollisella tavalla yleistys assosiaatiosäännöistä: esimerkiksi assosiaatiosääntöä

$$\{A, B, C\} \Rightarrow \{D, E, F\}$$

vastaa propositiosääntö

$$A \wedge B \wedge C \Rightarrow D \wedge E \wedge F.$$

Algoritmi B sopii siis sellaisenaan puhtaasti konjunktiivisten propositiosääntöjen etsimiseen, mutta se ei yleisty suoraan monimutkaisemmille säännöille.

Yksi selvä ongelma on, että kaavoilla ei ole lauseen 4 tapaista monotonisuusominaisuutta. Siksi algoritmia A ei voi ainakaan sellaisenaan käyttää kaavoille, joten täytyy tutkia muita lähestymistapoja. Yksi mahdollisuus on pyrkiä laskemaan kaavan kattavuus käyttämällä apuna hahmoja. Seuraava lause auttaa tässä.

Lause 6. *Jos φ on skeeman S kaava, on olemassa sellaiset hahmot $T_0, \ldots, T_{p-1} \subseteq S$ ja kertoimet $e_0, \ldots, e_{p-1} \in \mathbb{Z}$, että kaikilla S:n relaatioilla R pätee*

$$\mathrm{freq}_R(\varphi) = \sum_{j=0}^{p-1} e_j \, \mathrm{freq}_R(T_j). \tag{4}$$

Todistus. Lause 6 riittää todistaa relaatioille R, joiden koko on 1. Olkoot nimittäin luvut e_i ja hahmot T_i, $i \in [p]$, sellaiset, että yhtälö (4) on voimassa kaikille yksimonikkoisille relaatioille R. Kaavan tai hahmon kattavuus mielivaltaisessa relaatiossa R saadaan laskemalla yhteen sen kattavuus jokaisen monikon $v \in R$ yksin muodostamalle relaatiolle (v) ja jakamalla summa monikkojen lukumäärällä. Siis

$$\begin{aligned}
|R| \, \mathrm{freq}_R(\varphi) &= \sum \Big(\mathrm{freq}_{(v)}(\varphi) \;\Big|\; v \in R \Big) \\
&= \sum \Big(\sum_{j=0}^{p-1} e_j \, \mathrm{freq}_{(v)}(T_j) \;\Big|\; v \in R \Big) \\
&= \sum_{j=0}^{p-1} e_j \sum \Big(\mathrm{freq}_{(v)}(T_j) \;\Big|\; v \in R \Big) \\
&= \sum_{j=0}^{p-1} e_j \, |R| \, \mathrm{freq}_R(T_j).
\end{aligned}$$

Nyt siis riittää todistaa lauseen 6 se erikoistapaus, jossa relaatio R sisältää tasan yhden monikon v. Silloin jokaisen kaavan ja hahmon kattavuus on joko 0 tai 1. Käytetään induktiota kaavan φ rakenteen suhteen: määritelmästä 7 saadaan neljä tapausta.

Tapaus 1: φ on atomikaava, $\varphi = A$ jollain $A \in S$. Asetetaan $p = 1$, $e_1 = 1$ ja $T_1 = \{A\}$. Silloin $\mathrm{freq}_R(\varphi) = \mathrm{freq}_R(T_1)$, joten lause pätee φ:lle.

Tapaus 2: φ on negaatio, $\varphi = \neg\psi$. Olkoon kaavan ψ kattavuudella yhtälön (4) mukainen esitys

$$\mathrm{freq}_R(\psi) = \sum_{j=0}^{p'-1} e'_j \, \mathrm{freq}_R(T'_j).$$

Asetetaan $p = p' + 1$, $e_j = -e_j$ ja $T_j = T'_j$ kaikilla $j \in [p']$ sekä $e_{p'} = 1$ ja $T_{p'} = \emptyset$. Koska $\mathrm{freq}_R(\varphi) = 1 - \mathrm{freq}_R(\psi)$ ja $\mathrm{freq}_R(\emptyset) = 1$, kaava (4) pätee.

Tapaus 3: φ on konjunktio, $\varphi = (\psi \wedge \theta)$. Olkoot e'_i, T'_i, e''_j, T''_j ($i \in [p']$, $j \in [p'']$) sellaiset kokonaisluvut ja hahmot, että

$$\mathrm{freq}_R(\psi) = \sum_{j=0}^{p'-1} e'_i \, \mathrm{freq}_R(T'_i) \qquad \text{ja} \qquad \mathrm{freq}_R(\theta) = \sum_{j=0}^{p''-1} e''_j \, \mathrm{freq}_R(T''_j).$$

Asetetaan $p = p'p''$, $e_{p'j+i} = e'_i e''_j$ ja $T_{p'j+i} = T'_i \cup T''_j$ kaikilla $i \in [p']$, $j \in [p'']$.

Kirjoitetaan summa (4) muotoon

$$F = \sum_{j=0}^{p-1} e_j \operatorname{freq}_R(T_j) = \sum_{j=0}^{p''-1} F_j, \tag{5}$$

missä

$$F_j = \sum_{i=0}^{p'-1} e_{p'j+i} \operatorname{freq}_R(T_{p'j+i}) = e''_j \sum_{i=0}^{p'-1} e'_i \operatorname{freq}_R(T'_i \cup T''_j).$$

Jos $\operatorname{freq}_R(T''_j) = 0$, kaikilla T'_i pätee $\operatorname{freq}_R(T'_i \cup T''_j) = 0$, joten $F_j = 0$.

Jos $\operatorname{freq}_R(T''_j) = 1$, hahmo T''_j on monikon v osajoukko, joten $\operatorname{freq}_R(T'_i \cup T''_j) = \operatorname{freq}_R(T'_i)$ kaikilla $i \in [p']$. Silloin $F_j = e''_j \operatorname{freq}_R(\psi)$. Jos $v \models \psi$, saadaan $F_j = e''_j$, ja jos $v \not\models \psi$, saadaan $F_j = 0$.

Nyt siis

$$F_j = \begin{cases} 0 & \text{jos } \operatorname{freq}_R(T''_j) = 0 \text{ tai } \operatorname{freq}_R(\psi) = 0 \\ e''_j & \text{jos } \operatorname{freq}_R(T''_j) = 1 \text{ ja } \operatorname{freq}_R(\psi) = 1 \end{cases}$$

eli $F_j = e''_j \operatorname{freq}_R(T''_j) \operatorname{freq}_R(\psi)$. Ottamalla $\operatorname{freq}_R(\psi)$ yhteiseksi tekijäksi saadaan

$$F = \sum_{j=0}^{p''-1} F_j = \operatorname{freq}_R(\psi) \sum_{j=0}^{p''-1} e''_j \operatorname{freq}_R(T''_j) = \operatorname{freq}_R(\psi) \operatorname{freq}_R(\theta).$$

Yhden monikon relaatiossa R pätee $\operatorname{freq}_R(\varphi) = \operatorname{freq}_R(\psi \wedge \theta) = \operatorname{freq}_R(\psi) \operatorname{freq}_R(\theta)$, joten luvut e_j ja hahmot T_j, $j \in [p]$, toteuttavat yhtälön (4).

Tapaus 4: φ on disjunktio, $\varphi = (\psi \vee \theta)$. Kaava $\neg(\neg\psi \wedge \neg\theta)$ on de Morganin lakien perusteella ekvivalentti φ:n kanssa, joten tapaus 4 palautuu tapauksiin 2 ja 3. □

Huomautus 2. Konstruktion tuottamassa jonossa $(\, T_i \mid i \in [p]\,)$ voi sama hahmo toistua useita kertoja. Toistettuja hahmoja vastaavat termit kannattaa käytännössä tietysti yhdistää.

Huomautus 3. Tätä lausetta vastaa lähteen [MT96] väite "Proposition 1". Sen mukaan jokainen kerroin e_i voidaan aina valita joukosta $\{-1, +1\}$. Jos hahmojen T_i toistaminen sallitaan, tämä on ilmiselvästi mahdollista: edellinen konstruktio ilman huomautuksen 1 yhdistämisvaihetta antaa juuri sellaisen summan. Jos vaaditaan, että $T_i \neq T_j$, kun $i \neq j$, väite ei pidä paikkaansa. Tämä nähdään asettamalla $S = \{A, B\}$ ja $\varphi = (A \wedge B) \vee (\neg A \wedge \neg B)$. Edellisestä todistuksesta saadaan tulos $\operatorname{freq}_R(\varphi) = \operatorname{freq}_R(\emptyset) - \operatorname{freq}_R(\{A\}) - \operatorname{freq}_R(\{B\}) + 2 \operatorname{freq}_R(\{A, B\})$.

Jos relaatiossa R on 20 monikkoa ja hahmojen kattavuudet ovat $\operatorname{freq}_R(\{A\}) = 10/20$, $\operatorname{freq}_R(\{B\}) = 15/20$ ja $\operatorname{freq}_R(\{A, B\}) = 6/20$, saadaan $\operatorname{freq}_R(\varphi) = 7/20$. Olkoon F kaavan (4) tyyppinen summa, jossa hahmot T_i eivät toistu ja kertoimet $e_i \in \{-1, +1\}$. Tarkastellaan lukua $N = 20F$. Jos mikään hahmoista T_i ei ole $\{A, B\}$, N on viidellä jaollinen, sillä jokaiselle hahmolle $T \in \{\emptyset, \{A\}, \{B\}\}$ pätee, että $20 \operatorname{freq}_R(T)$ on viidellä jaollinen. Jos taas jokin hahmoista T_i on $\{A, B\}$, joko $N - 6$ tai $N + 6$ on viidellä jaollinen. Missään tapauksessa ei voi olla $N = 7$.

Ongelma. Kaavan kattavuudesta saadaan taas helposti säännön tuki ja luotettavuus, tosin algoritmi B ei siltikään ole käyttökelpoinen sellaisenaan. Yksi mahdollinen tavoite on algoritmi, joka etsii ensin ρ-kattavat assosiaatiosäännöt ja yhdistelee niistä sitten jonkin kiinnostavuuskriteerin avulla yleisempiä sääntöjä. Näiden sääntöjen luotettavuus on tietysti helppo laskea vielä yhdellä tietokannan läpikäynnillä. Sääntöjä tulee kuitenkin helposti eksponentiaalisen paljon, ja suuri osa voitaisiin todennäköisesti karsia jo laskettujen kattavien joukkojen avulla.

Rajoitutaan yksinkertaisuuden vuoksi tarkastelemaan sääntöä

$$\varphi \quad = \quad B \Rightarrow A_0 \vee A_1 \vee \cdots \vee A_{n-1}.$$

Sen luotettavuuden määrittelee kaava (3). Nimittäjässä esiintyvä kattavuus $\mathrm{freq}(B)$ tunnetaan, koska algoritmia A käytettäessä saadaan joka tapauksessa yksialkioisten hahmojen kattavuudet. Osoittajassa on kaavan

$$B \wedge (A_0 \vee \cdots \vee A_{n-1})$$

kattavuus, joka saadaan joko edellisen todistuksen konstruktiosta tai lauseen 7 (sivulla 19) avulla inkluusio-ekskluusio -summana

$$\mathrm{freq}(\varphi) = \sum \left((-1)^{|J|+1} \,\mathrm{freq}(A_J) \;\middle|\; \emptyset \neq J \subseteq [n] \right),$$

missä $A_J = \{\, A_j \mid j \in J \,\} \cup \{B\}$. Summasta ei kuitenkaan välttämättä tunneta kaikkia termejä, mutta sen sijaan algoritmi voi helposti laskea summan

$$\mathrm{freq}'(\varphi, \rho) = \sum \left((-1)^{|J|+1} \,\mathrm{freq}(A_J) \;\middle|\; \emptyset \neq J \subseteq [n],\, \mathrm{freq}(A_J) \geq \rho \right).$$

Ongelma on nyt selvittää, kuinka suuri virhe tehdään, toisin sanoen arvioida lukua

$$|\,\mathrm{freq}(\varphi) - \mathrm{freq}'(\varphi, \rho)|.$$

Kattavuuden monotonisuuden ansiosta summa $\mathrm{freq}'(\varphi, \rho)$ lasketaan siinä mielessä alaspäin suljetun joukon suhteen, että jos $\mathrm{freq}(A_J)$ on mukana summassa, siinä on myös $\mathrm{freq}(A_{J'})$ kaikilla $J' \subseteq J$. Käytettävissä on myös algoritmin A sivutuotteena tieto minimaalisten ei-kattavien joukkojen kattavuuksista.

Tämän tilanteen erikoistapauksena voisi käydä niin, että ρ-kattavat hahmot olisivat täsmälleen skeeman S enintään k-alkioiset osajoukot jollakin $k \in \mathbb{N}$. Tässä tapauksessa minimaaliset ei-kattavat hahmot ovat S:n $k+1$-alkioiset osajoukot, ja voidaan soveltaa Bonferronin epäyhtälöä (lause 11 sivulla 28). Sen mukaan virhe on enintään minimaalisten ei-kattavien hahmojen kattavuuksien summa. Arvioimalla kattavuuksia ylöspäin saadaan virheelle raja $N\rho$, missä N on $k+1$-alkioisten hahmojen lukumäärä ja ρ kattavuusraja.

Yleisessä tapauksessa tarvittaisiin jokin muu ratkaisu. Seuraavat luvut esittelevät matemaattisia välineitä, ja luvussa 6 palataan tähän ongelmaan.

4 Inkluusio-ekskluusio: vanhoja tuloksia

Tässä luvussa esitetään klassinen inkluusio-ekskluusio -periaate ja joitakin sen alkeellisia yleistyksiä. Inkluusio-ekskluusio on yksinkertaisen tuntuinen idea; kirjan [Knu73] (luku 1.3.3) mukaan sen on keksinyt de Moivre 1700-luvulla, mutta se on tullut yleisesti tunnetuksi vasta Whitworthin kirjoitettua siitä vuonna 1867. Bonferroni-tyyppiset epäyhtälöt ovat kirjan [GS96] mukaan peräisin Boolen ja Jordanin töistä vuosilta 1854 ja 1927.

Sopimus. Merkitköön luvuissa 4, 5 ja 6 U jotain kiinnitettyä äärellistä perusjoukkoa ("universumia") ja $\mathcal{A}$ kiinnitettyä n-alkioista jonoa U:n osajoukkoja, $\mathcal{A} = (\, A_j \mid j \in [n] \,)$. Tavoite on laskea yhdisteen $\bigcup \mathcal{A}$ koko tai arvioida sitä.

4.1 Inkluusio-ekskluusio

Lause 7. *Jonon $\mathcal{A}$ joukkojen yhdisteen $\bigcup \mathcal{A}$ koko voidaan laskea kaavalla*

$$|\bigcup \mathcal{A}| = \sum \Big((-1)^{|\mathcal{B}|+1} \, |\bigcap \mathcal{B}| \;\Big|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A} \Big) \tag{6}$$

kun tunnetaan jonon $\mathcal{A}$ joukkojen kaikkien leikkausten koot.

Lause 7 tunnetaan inkluusio-ekskluusio -periaatteena tai summan ja erotuksen periaatteena. Nimityksen "inkluusio-ekskluusio" taustalla lienee seuraava ajattelu: Lasketaan ensin yhteen kaikkien joukkojen $X \in \mathcal{A}$ koot. Silloin ne alkiot, jotka kuuluvat kahteen joukkoon, tulevat lasketuiksi vähintään kaksi kertaa, joten vähennetään joukkojen $X, Y \in \mathcal{A}$, $X \neq Y$, parittaisten leikkausten koot $|X \cap Y|$. Nyt kuitenkin vähennettiin liian monta kertaa ne alkiot, jotka kuuluvat kolmeen joukkoon, joten lisätään kolmittaisten leikkausten koot $|X \cap Y \cap Z|$, $X, Y, Z \in \mathcal{A}$, $X \neq Y \neq Z \neq X$. Neljään joukkoon kuuluvat alkiot lisättiin liian monta kertaa, ja niin edelleen.

Tästä ideasta saadaan pitävä todistus, kun tarkastellaan erikseen jokaisen alkion vaikutusta summaan. Asetetaan sitä varten ensin seuraava määritelmä.

Määritelmä 11. Kun $x \in U$,

$$\mathcal{A}_x = (\, X \in \mathcal{A} \mid x \in X \,)$$

on niiden $\mathcal{A}$:n joukkojen jono, jotka sisältävät alkion x.

Lauseen 7 todistus. Kirjoitetaan yhtälön (6) vasen puoli summaksi

$$\sum_{x \in \bigcup \mathcal{A}} 1$$

ja vastaavasti oikea puoli kaksinkertaiseksi summaksi

$$\sum \Big((-1)^{|\mathcal{B}|+1} \sum_{x \in \bigcap \mathcal{B}} 1 \;\Big|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A} \Big).$$

Oikean puolen sisempi summa lasketaan joidenkin U:n alkioiden suhteen, samoin kuin vasemman puolen summa. Vaihdetaan siis summausjärjestystä. Kun $\mathcal{B} \subseteq \mathcal{A}$, ehto $x \in \bigcap \mathcal{B}$ voidaan kirjoittaa muotoon $\mathcal{B} \subseteq \mathcal{A}_x$, joten oikea puoli on

$$\sum \left((-1)^{|\mathcal{B}|+1} \;\middle|\; x \in U, \emptyset \neq \mathcal{B} \subseteq \mathcal{A}, \mathcal{B} \subseteq \mathcal{A}_x \right) = \sum_{x \in U} \sum \left((-1)^{|\mathcal{B}|+1} \;\middle|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A}_x \right).$$

Sisempi summa on selvästi 0, jos $x \notin \bigcup \mathcal{A}$. Väite seuraa, jos voidaan osoittaa, että epätyhjille äärellisille joukoille X pätee kaava

$$\sum \left((-1)^{|Y|+1} \;\middle|\; \emptyset \neq Y \subseteq X \right) = 1$$

tai yhtäpitävästi

$$\sum \left((-1)^{|Y|} \;\middle|\; Y \subseteq X \right) = 0. \tag{7}$$

Tähän voidaan nimittäin sijoittaa joukoksi X jonon $\mathcal{A}_x$ indeksijoukko.

Jos $|X| = m$ ja $j \in \mathbb{Z}$, joukolla X on $\binom{m}{j}$ j-osajoukkoa. Siis yhtälön (7) vasen puoli voidaan kirjoittaa muotoon

$$\begin{aligned}\sum_{Y \subseteq X} (-1)^{|Y|} &= \sum_{j=0}^{m} \sum \left((-1)^j \;\middle|\; Y \in X^{(j)} \right) \\ &= \sum_{j=0}^{m} (-1)^j \sum \left(1 \;\middle|\; Y \in X^{(j)} \right) = \sum_{j=0}^{m} (-1)^j \binom{m}{j}.\end{aligned}$$

Binomilauseen (lause 3) mukaan $\sum_{j=0}^{m} \binom{m}{j} (-1)^j 1^{m-j} = (1-1)^m = 0$, kun $m > 0$, joten väite on todistettu. □

Esimerkki 2. Olkoot $p_0 = 2$, $p_1 = 3$, $p_2 = 5$ ja $p_3 = 7$. Olkoon $U = \{1, 2, 3, \ldots, 100\}$ ja $\mathcal{A} = (A_0, A_1, A_2, A_3)$, missä

$$A_j = \{ \, x \in U \mid x \text{ on jaollinen luvulla } p_j \, \}.$$

Toisin sanottuna $A_0 = \{2, 4, \ldots, 98, 100\}$, $A_1 = \{3, 6, \ldots, 96, 99\}$, $A_2 = \{5, 10, \ldots, 95, 100\}$ ja $A_3 = \{7, 14, \ldots, 91, 98\}$. Koska luvut p_j ovat alkulukuja, joukkojen A_j leikkauksille pätee

$$\bigcap_{j \in J} A_j = \{ \, x \in U \mid x \text{ on jaollinen luvulla } \prod_{j \in J} p_j \, \},$$

missä $J \subseteq [4]$. Helposti nähdään, että tällaisen leikkauksen koko on

$$\left| \bigcap_{j \in J} A_j \right| = \left\lfloor 100 / \prod_{j \in J} p_j \right\rfloor,$$

missä $\lfloor \cdot \rfloor$ tarkoittaa alaspäin pyöristystä. Siis leikkausten ko'oiksi saadaan seuraavat:

$$|A_0| = 50,\ |A_1| = 33,\ |A_2| = 20,\ |A_3| = 14,\ |A_0 \cap A_1| = 16,\ |A_0 \cap A_2| = 10,\ |A_0 \cap A_3| = 7,$$
$$|A_1 \cap A_2| = 6,\ |A_1 \cap A_3| = 4,\ |A_2 \cap A_3| = 2,\ |A_0 \cap A_1 \cap A_2| = 3,\ |A_0 \cap A_1 \cap A_3| = 2,$$
$$|A_0 \cap A_2 \cap A_3| = 1,\ |A_1 \cap A_2 \cap A_3| = 0,\ |A_0 \cap A_1 \cap A_2 \cap A_3| = 0.$$

Nyt lauseesta 7 seuraa, että joukossa U on neljällä ensimmäisellä alkuluvulla jaollisia lukuja

$$|\bigcup \mathcal{A}| = 50 + 33 + 20 + 14 - 16 - 10 - 7 - 6 - 4 - 2 + 3 + 2 + 1 + 0 - 0 = 78.$$

Esimerkki 3. Kuva 1 esittää jonon $\mathcal{A} = (A, B, C, D)$ osajonohilaa, jossa on lyhyyden vuoksi lueteltu osajonot ilman tavanomaisia sulkeita ja pilkkuja. Nuolet kuvaavat osajonorelaatiota; tarkemmin sanottuna joukon $\mathcal{P}(\mathcal{A})$ osajonorelaatio on nuolten esittämän relaation refleksiivinen ja transitiivinen sulkeuma.

Inkluusio-ekskluusio -summassa (6) on tämän esimerkin tapauksessa 15 termiä: jokaista kuvan 1 epätyhjää joukkoa vastaa termi. Summan voi muodostaa kuvan avulla: lasketaan jokaisella vaakarivillä lueteltujen jonojen leikkausten koot yhteen ja kootaan tuloksista vuorotteleva summa.

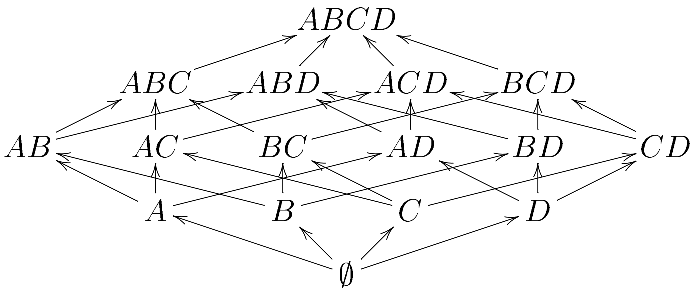
Esimerkin ajatus tasoittaisista summista on hyödyllinen yleistettäessä tulosta, joten asetetaan seuraava määritelmä.

Määritelmä 12. Kun $j \in \mathbb{Z}^+$, perheen $\mathcal{A}$ j:s tasoittainen summa on

$$S_j = \sum \Big(|\bigcap \mathcal{B}| \ \Big| \ \mathcal{B} \in \mathcal{A}^{(j)} \Big).$$

Selvästikin summa S_j eroaa nollasta vain, jos $j \leq n$.

Lauseen 7 todistuksessa oli hyödyllistä kirjoittaa suureita summina universumin alkioiden suhteen. Myös tasoittaiset summat voidaan muotoilla tällä tavalla.



Kuva 1: Jonon (A, B, C, D) osajonohila

Lemma 8. *Olkoon $j \in \mathbb{N}$. Kun $x \in U$, merkitään $s_x = |\mathcal{A}_x|$. Silloin*

$$S_j = \sum_{x \in U} \binom{s_x}{j}.$$

Todistus. Kirjoitetaan S_j:n määritelmässä esiintyvä joukon $\bigcap \mathcal{B}$ koko vakion 1 summana sen alkioiden suhteen ja yhdistetään tämä ulompaan summaan.

$$S_j = \sum \Big(\sum (1 \mid x \in \bigcap \mathcal{B}) \Bigm| \mathcal{B} \in \mathcal{A}^{(j)} \Big) = \sum \Big(1 \Bigm| \mathcal{B} \in \mathcal{A}^{(j)},\, x \in \bigcap \mathcal{B} \Big).$$

Ehto $x \in \bigcap \mathcal{B}$ on yhtäpitävä sen kanssa, että $\mathcal{B} \subseteq \mathcal{A}_x$, joten summa saadaan muotoon

$$S_j = \sum \Big(1 \Bigm| x \in U,\, \mathcal{B} \in \mathcal{A}_x^{(j)} \Big) = \sum_{x \in U} |\mathcal{A}_x^{(j)}| = \sum_{x \in U} \binom{s_x}{j},$$

mikä todistaa väitteen. □

Tasoittaisten summien avulla voidaan välittömästi kirjoittaa inkluusio-ekskluusio -väite muotoon

$$|\bigcup \mathcal{A}| = \sum_{j \geq 1} (-1)^{j+1} S_j = \sum_{j=1}^{n} (-1)^{j+1} S_j. \tag{8}$$

Tämä muotoilu on siinä mielessä helpompi hahmottaa kuin yhtälö (6), että summaus tehdään numeroitujen tasojen eikä suoraan kaikkien joukkojen suhteen.

Esimerkki 4. Olkoon $U = [10] = \{0, 1, \ldots, 9\}$ ja $\mathcal{A} = (A_0, \ldots, A_4)$, missä joukot A_j on kuvattu taulukossa 4. Taulukosta nähdään heti, että $|\bigcup \mathcal{A}| = 8$, mutta suureiden S_j lukeminen on työläämpää. Tuloksiksi saadaan

$$\begin{aligned}
S_1 &= |A_0| + |A_1| + |A_2| + |A_3| + |A_4| = 18, \\
S_2 &= |A_0 \cap A_1| + |A_0 \cap A_2| + \cdots + |A_3 \cap A_4| = 16, \\
S_3 &= |A_0 \cap A_1 \cap A_2| + \cdots + |A_2 \cap A_3 \cap A_4| = 8, \\
S_4 &= |A_0 \cap A_1 \cap A_2 \cap A_3| + \cdots + |A_1 \cap A_2 \cap A_3 \cap A_4| = 2, \\
S_5 &= |A_0 \cap A_1 \cap A_2 \cap A_3 \cap A_4| = 0
\end{aligned}$$

ja siis $|\bigcup \mathcal{A}| = 18 - 16 + 8 - 2 + 0 = 8$.

Tässä tapauksessa inkluusio-ekskluusio -periaatteen soveltaminen ei ole kovin järkevää, koska taulukosta on helpompi lukea lopputulos kuin välivaiheet. Seuraavan esimerkin tapauksessa tilanne on päinvastainen.

Esimerkki 5. Tutkitaan äärellisen joukon X permutaatioita eli bijektioita $p : X \to X$. Yleisyyttä rajoittamatta voidaan olettaa, että $X = [m]$ jollain $m \in \mathbb{N}$. Merkitään tämän joukon permutaatioiden joukkoa U:lla; tunnetun lauseen mukaan $|U| = m!$. Alkion $x \in X$

	0	1	2	3	4	5	6	7	8	9
A_0	•	•	•			•				•
A_1	•			•			•			
A_2			•		•					
A_3	•	•	•		•					
A_4	•		•	•		•				

Taulukko 3: Esimerkkiin 4 liittyvä taulukko

sanotaan olevan permutaation (tai minkä tahansa kuvauksen) $p : X \to X$ *kiintopiste*, jos $p(x) = x$. Tehtävä on laskea, kuinka monella permutaatiolla ei ole yhtään kiintopistettä.

Inkluusio-ekskluusion käyttämiseksi asetetaan $\mathcal{A} = (\, A_i \mid i \in [m]\,)$, missä $A_i = \{\, p \in U \mid p(i) = i \,\}$ on niiden permutaatioiden joukko, joille i on kiintopiste. Jos nyt $\mathcal{B} \subseteq \mathcal{A}$, $\mathcal{B} = (A_{i_1}, A_{i_2}, \ldots, A_{i_j})$, perheeseen $\bigcap \mathcal{B}$ kuuluvat täsmälleen ne permutaatiot $p \in U$, joille $i_1, i_2, \ldots, i_j$ ovat kiintopisteitä. Sellaisia on yhtä monta kuin jäljelle jäävien $m - j$:n alkion permutaatioita eli $(m-j)!$. Toisaalta tällaisia jonoja $\mathcal{B}$ on $\binom{m}{j}$. Siis määritelmän 12 tasoittainen summa on $S_j = \binom{m}{j}(m-j)! = \frac{m!}{j!}$, kun $0 \le j \le m$.

Kiintopisteettömiä permutaatioita on nyt inkluusio-ekskluusio -periaatteen (lauseen 7) perusteella

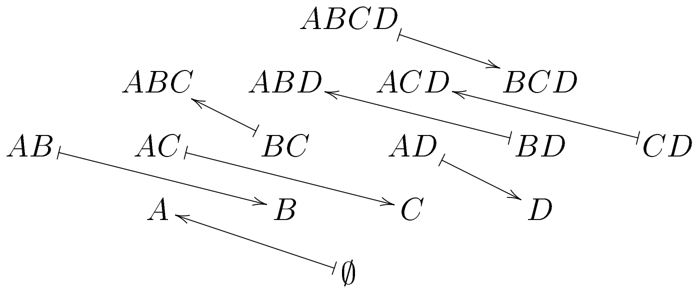
$$m! - |\bigcup \mathcal{A}| = m! \left(1 - \sum_{j=1}^{m} (-1)^{j+1} \frac{m!}{j!}\right).$$

Suluissa oleva lauseke, kiintopisteettömien permutaatioiden osuus, muistuttaa funktion e^x Maclaurinin sarjaa $\sum_j x^j/j!$. Kyseessä on itse asiassa m:s Maclaurinin polynomi laskettuna pisteessä -1. Taylorin kaavan virhetermi on tässä tapauksessa $(-1)^{m+1}e^{\xi}/(m+1)!$, missä $-1 \le \xi \le 0$, joten polynomi lähestyy lukua $1/e$, kun $m \to \infty$.

Palataan vielä lauseen 7 todistuksessa esiintyneeseen yhtälöön (7), jonka olennainen sisältö on, että epätyhjällä äärellisellä joukolla on yhtä monta pariton- ja parillisalkioista osajoukkoa. Väite todistettiin binomilauseen avulla, ja se seuraa myös lemmasta 9 (sivulla 25). Niin alkeellisen tuloksen voi kuitenkin todistaa alkeellisella menetelmällä, ja seuraava todistus on kiinnostava lähinnä menetelmän takia.

Merkitään $\mathcal{E} = \{\, Y \subseteq X \mid \#Y \text{ parillinen} \,\} = \bigcup\{\, X^{(2j)} \mid j \in \mathbb{N} \,\}$ ja $\mathcal{O} = \{\, Y \subseteq X \mid \#Y \text{ pariton} \,\} = \bigcup\{\, X^{(2j+1)} \mid j \in \mathbb{N} \,\}$. Valitaan $a \in X$ ja määritellään kuvaus $f : \mathcal{E} \to \mathcal{O}$ asettamalla

$$f(Y) = Y \mathbin{\triangle} \{a\} \text{ kaikilla } Y \in \mathcal{E}. \tag{9}$$



Kuva 2: Eräs bijektio $\mathcal{E} \to \mathcal{O}$ tapauksessa $X = \{A, B, C, D\}$

Tässä $\triangle$ merkitsee symmetristä erotusta. Jos $a \in Y$, $f(Y) = Y \setminus \{a\}$ ja $|f(Y)| = |Y \setminus \{a\}| = |Y| - 1$. Jos $a \notin Y$, $f(Y) = Y \cup \{a\}$ ja $|f(Y)| = |Y \cup \{a\}| = |Y| + 1$. Kummassakin tapauksessa $f(Y) \in \mathcal{O}$, joten f on hyvin määritelty. Jos taas $Y \in \mathcal{O}$, vastaavalla tavalla nähdään, että $Y \triangle \{a\} \in \mathcal{E}$. Koska edelleen $f(Y \triangle \{a\}) = Y$, kuvaus f on surjektio. Jos $f(Y) = f(Y')$ eli $Y \triangle \{a\} = Y' \triangle \{a\}$, on välttämättä $Y = Y'$. Siis f on bijektio, ja joukot $\mathcal{E}$ ja $\mathcal{O}$ on osoitettu samankokoisiksi.

Myöhempää varten on hyvä huomata, että yhtälön (9) määrittelemä bijektio f kuvaa jokaisen joukon $Y \in \mathcal{E}$ joko yhtä alkiota suuremmaksi tai yhtä alkiota pienemmäksi joukoksi. Esimerkki tästä bijektiosta tapauksessa $X = \{A, B, C, D\}$ annetaan kuvassa 2.

4.2 Inkluusio-ekskluusion yleistys

Lauseen 7 voi tulkita myös probabilistisesti. Nimitetään U:n osajoukkoja *tapahtumiksi* ja kirjoitetaan $\mathsf{P}(X) = |X| \,/\, |U|$ tapahtumilla $X \subseteq U$. Silloin todennäköisyys, että vähintään yksi $\mathcal{A}$:n tapahtumista toteutuu, on

$$\mathsf{P}\left(\bigcup \mathcal{A}\right) = \sum \Big((-1)^{|\mathcal{B}|+1} \mathsf{P}\left(\bigcap \mathcal{B}\right) \;\Big|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A} \Big). \tag{10}$$

Tässä tulkinnassa määritelmän 11 jono $\mathcal{A}_x$ luettelee ne $\mathcal{A}$:n tapahtumat, jotka toteutuvat alkeistapauksessa x.

Edellä on vain kirjoitettu kombinatorinen lause eri tavalla, mutta kappaleessa 4.4 esitetään, miten (10) ja muita tuloksia yleistetään mielivaltaisiin todennäköisyysavaruuksiin. Seuraavassa keskitytään etsimään vastausta todennäköisyystulkinnasta kumpuaviin kysymyksiin siitä, millä todennäköisyydellä jokin määrä jonon $\mathcal{A}$ tapahtumista toteutuu. Kysymysten tarkempaa muotoilua varten asetetaan seuraava määritelmä.

Määritelmä 13. Olkoon $k \in \mathbb{N}$. Joukko $P_k \subseteq U$ sisältää ne alkiot x, jotka kuuluvat *vähintään* k:hon jonon $\mathcal{A}$ joukoista,

$$P_k = \{\, x \in U \mid s_x \geq k \,\}.$$

Vastaavasti joukko Q_k sisältää ne alkiot x, jotka kuuluvat *täsmälleen* k:hon jonon $\mathcal{A}$ joukoista,

$$Q_k = \{\, x \in U \mid s_x = k \,\}.$$

Tässä $s_x = |\mathcal{A}_x|$, kuten lemmassa 8.

Inkluusio-ekskluusio -periaate (lause 7) vastaa kysymykseen, kuinka moni alkio kuuluu vähintään yhteen $\mathcal{A}$:n joukkoon, ts. mikä on joukon P_1 koko eli millä todennäköisyydellä jonon $\mathcal{A}$ tapahtumista toteutuu vähintään yksi. Seuraavassa yleistetään vastaus kaikille P_k ja Q_k.

Tarkastellaan joukon Q_k koon laskemista epämuodollisesti samaan tapaan kuin lausetta 7 perusteltiin. Etsitään vastausta muodossa

$$|Q_k| = \sum_{j=k}^{n} q_j S_j,$$

missä lukujono $(\, q_j \mid j \geq k)$ riippuu vain k:sta.

Ilmeisesti on valittava $q_k = 1$, jotta taso $\mathcal{A}^{(k)}$ tulee lasketuksi mukaan. Silloin tasoon $\mathcal{A}^{(k+1)}$ liittyvät leikkaukset lasketaan $k+1$ kertaa, koska jokaiseen $k+1$-joukkoon sisältyy $k+1$ eri k-joukkoa. Asetetaan $q_{k+1} = -(k+1)$, koska tätä tasoa ei haluta laskea lainkaan. Tason $\mathcal{A}^{(k+2)}$ joukot on laskettu ensin $\binom{k+2}{k} = (k+1)(k+2)/2$ kertaa, mutta poistettu sitten $(k+1)\binom{k+2}{k+1} = (k+1)(k+2)$ kertaa. Asetetaan $q_{k+2} = (k+1)(k+2)/2$. Näin jatkamalla voidaan arvata ratkaisuksi $q_j = (-1)^{j-k}\binom{j}{k}$.

Koska $P_k = \bigcup_{j=k}^{n} Q_j$ on joukon P_k ositus, joukon P_k koko saadaan laskemalla yhteen lukuja $|Q_j|$. Laskussa esiintyy binomikerrointen vuorotteleva summa, jonka eräs erikoistapaus sisältyi myös lauseen 7 todistukseen (yhtälö (7) sivulla 20). Seuraavaa lemmaa käytetään myöhemmin myös Bonferronin epäyhtälön todistamiseen.

Lemma 9. *Olkoot $k, m \in \mathbb{N}$. Silloin*

$$\sum_{j=0}^{m} (-1)^j \binom{k}{j} = (-1)^m \binom{k-1}{m}.$$

Todistus. Tapaus $k=0$ on helppo: vasemman puolen summan ainoa nollasta eroava termi on $(-1)^0\binom{0}{0} = 1$, ja oikea puoli on $(-1)^m\binom{-1}{m} = 1$. Kiinnitetään $k>0$ ja todistetaan lause induktiolla m:n suhteen.

Tapauksessa $m=0$ väite seuraa lauseesta 1(a). Oletetaan, että väite on tosi jollain $m \in \mathbb{N}$ ja pyritään osoittamaan se $m+1$:lle. Tällöin summa on

$$\sum_{j=0}^{m+1} (-1)^j \binom{k}{j} = (-1)^m \binom{k-1}{m} + (-1)^{m+1}\binom{k}{m+1}$$
$$= (-1)^m \left[\binom{k-1}{m} - \binom{k}{m+1}\right] = (-1)^{m+1}\binom{k-1}{m+1},$$

missä viimeinen yhtäsuuruus seuraa lauseesta 1(d). □

Lause 10. *Olkoon $0 < k \le n$ (missä $n = |\mathcal{A}|$). Joukoille P_k ja Q_k pätee*

$$|P_k| = \sum_{j=k}^{n} (-1)^{j-k} \binom{j-1}{k-1} S_j \tag{11}$$

ja

$$|Q_k| = \sum_{j=k}^{n} (-1)^{j-k} \binom{j}{k} S_j. \tag{12}$$

Todistus. Aloitetaan yhtälöstä (12). Kirjoitetaan sen oikea puoli summaksi joukon U alkioiden suhteen. Lemman 8 avulla saadaan

$$\begin{aligned}
\sum_{j=k}^{n} (-1)^{j-k} \binom{j}{k} S_j &= \sum_{x \in U} \sum_{j=k}^{n} (-1)^{j-k} \binom{j}{k} \binom{s_x}{j} \\
&= \sum_{x \in U} \sum_{j=k}^{n} (-1)^{j-k} \binom{s_x}{k} \binom{s_x - k}{j - k} \\
&= \sum_{x \in U} \binom{s_x}{k} \sum_{j=k}^{n} (-1)^{j-k} \binom{s_x - k}{j - k}.
\end{aligned}$$

Toinen yhtäsuuruus seuraa lauseesta 1(f). Niitä alkioita x vastaavat termit, joille $s_x < k$, ovat selvästi nollia. Jos taas jollain x on $s_x > k$, sisemmän summan nähdään jälleen binomilauseen (lause 3) avulla olevan binomin potenssi $(1-1)^{s_x - k} = 0$. Kaikki tarvittavat termit ovat summassa mukana, sillä $s_x \le n$ ja siis $s_x - k \le n - k$ millä tahansa x.

Jäljelle jäävät ne alkiot $x \in U$, joille $s_x = k$. Tässä tapauksessa $\binom{s_x}{k} = 1$ ja sisemmän summan ainoa nollasta poikkeava termi on $(-1)^{k-k} \binom{s_x - k}{k - k} = (-1)^0 \binom{0}{0} = 1$. Siis summa laskee täsmälleen joukon $Q_k = \{ x \in U \mid s_x = k \}$ alkiot.

Yhtälön (11) todistamiseksi havaitaan ensin, että joukolla P_k on ositus

$$P_k = \bigcup_{j=k}^{n} Q_j$$

parittain erillisiin joukkoihin Q_j. Siis $|P_k|$ voidaan laskea summana

$$|P_k| = \sum_{\ell=k}^{n} |Q_\ell| = \sum_{\ell=k}^{n} \sum_{j=\ell}^{n} (-1)^{j-\ell} \binom{j}{\ell} S_j.$$

Vaihtamalla summausjärjestystä saadaan

$$\begin{aligned}
|P_k| &= \sum_{j=k}^{n} S_j \sum_{\ell=k}^{j} (-1)^{j-\ell} \binom{j}{\ell} \\
&= \sum_{j=k}^{n} S_j (-1)^j \left(\sum_{\ell=0}^{j} (-1)^{\ell} \binom{j}{\ell} - \sum_{\ell=0}^{k-1} (-1)^{\ell} \binom{j}{\ell} \right) \\
&= \sum_{j=k}^{n} S_j (-1)^j \left((-1)^j \binom{j-1}{j} - (-1)^{k-1} \binom{j-1}{k-1} \right) \\
&= \sum_{j=k}^{n} (-1)^{j+k} \binom{j-1}{k-1} S_j.
\end{aligned}$$

Toiseksi viimeinen yhtäsuuruus seuraa lemmasta 9 ja viimeinen lauseesta 1(b). Koska $(-1)^{j+k} = (-1)^{j-k}$, yhtälö (11) on todistettu. $\square$

Esimerkki 6. Jatketaan esimerkkiä 4. Lauseen 10 mukaan joukoille P_j pätee

$$\begin{aligned}
|P_1| &= \binom{0}{0} \cdot 18 - \binom{1}{0} \cdot 16 + \binom{2}{0} \cdot 8 - \binom{3}{0} \cdot 2 + \binom{4}{0} \cdot 0 = 8, \\
|P_2| &= \binom{1}{1} \cdot 16 - \binom{2}{1} \cdot 8 + \binom{3}{1} \cdot 2 - \binom{4}{1} \cdot 0 = 6, \\
|P_3| &= \binom{2}{2} \cdot 8 - \binom{3}{2} \cdot 2 + \binom{4}{2} \cdot 0 = 2, \\
|P_4| &= \binom{3}{3} \cdot 2 - \binom{4}{3} \cdot 0 = 2 \text{ ja} \\
|P_5| &= \binom{4}{4} \cdot 0 = 0,
\end{aligned}$$

ja joukoille Q_j

$$\begin{aligned}
|Q_1| &= \binom{1}{1} \cdot 18 - \binom{2}{1} \cdot 16 + \binom{3}{1} \cdot 8 - \binom{4}{1} \cdot 2 + \binom{5}{1} \cdot 0 = 2, \\
|Q_2| &= \binom{2}{2} \cdot 16 - \binom{3}{2} \cdot 8 + \binom{4}{2} \cdot 2 - \binom{5}{2} \cdot 0 = 4, \\
|Q_3| &= \binom{3}{3} \cdot 8 - \binom{4}{3} \cdot 2 + \binom{5}{3} \cdot 0 = 8 - 8 = 0, \\
|Q_4| &= \binom{4}{4} \cdot 2 - \binom{5}{4} \cdot 0 = 2 \text{ ja} \\
|Q_5| &= \binom{5}{5} \cdot 0 = 0.
\end{aligned}$$

4.3 Bonferronin epäyhtälö

Inkluusio-ekskluusio -periaatetta käytettäessä tarvitaan jonon $\mathcal{A}$ kaikkien osajonojen leikkausten koot, siis 2^n lukua. Tämä on liian paljon käytännössä laskettavaksi, joten seuraavassa keskitytään approksimoimaan summalauseketta. Lauseen 7 todistuksessa oli hyödyllistä tarkastella vuorottelevaa summaa tasoittain, j:n kokoisia osajoukkoja kerrallaan. Jos summa katkaistaan tällaisen tason S_j jälkeen, Bonferronin epäyhtälön mukaan virhe on enintään seuraavan tason S_{j+1} suuruinen. Toisin sanoen osasummat ovat vuorotellen ylä- ja alarajoja.

Epäyhtälöstä tulee mieleen Leibnizin suppenemisehto: jos alternoivan sarjan termien itseisarvojen jono on vähenevä ja lähestyy äärettömässä nollaa, sarja suppenee, ja sen osasummat ovat vuorotellen ylä- ja alarajoja. Kyse ei kuitenkaan ole samasta asiasta: yksinkertaisen esimerkin saa asettamalla $\mathcal{A}$:n jokaiseksi joukoksi U:n. Silloin lemman 8 mukaan $S_j = \binom{n}{j}|U|$, eivätkä binomikertoimet muodosta vähenevää jonoa, jos $n \geq 2$.

Lause 11 (Bonferronin epäyhtälö). *Olkoon $0 \leq k \leq n$. Merkitään inkluusio-ekskluusio -summaa, jossa on vain k ensimmäistä termiä,*

$$b(\mathcal{A}, k) = \sum_{j=1}^{k} (-1)^{j+1} S_j.$$

Edelleen merkitään tämän summan virhettä

$$d(\mathcal{A}, k) = \Big| |\bigcup \mathcal{A}| - b(\mathcal{A}, k) \Big|.$$

Silloin

$$d(\mathcal{A}, k) \leq S_{k+1}. \tag{13}$$

Huomautus 4. Väitteen voi muotoilla yhtäpitävästi näin: jos k on parillinen,

$$b(\mathcal{A}, k) \leq |\bigcup \mathcal{A}| \leq b(\mathcal{A}, k+1). \tag{14}$$

Olkoon nimittäin k pienin parillinen luku, jolla epäyhtälö 14 ei pidä paikkaansa; helposti nähdään, että $k \geq 2$, sillä $b(\mathcal{A}, 0) = 0$ ja $b(\mathcal{A}, 1) \geq |\bigcup \mathcal{A}|$. Jos $b(\mathcal{A}, k) > |\bigcup \mathcal{A}|$, oletuksen perusteella $b(\mathcal{A}, k-1) \geq |\bigcup \mathcal{A}|$, mutta koska $b(\mathcal{A}, k) = b(\mathcal{A}, k-1) - S_k$, saadaan $d(\mathcal{A}, k-1) = (b(\mathcal{A}, k-1) - b(\mathcal{A}, k)) + (b(\mathcal{A}, k) - |\bigcup \mathcal{A}|) > S_k$. Jos taas $b(\mathcal{A}, k) \leq |\bigcup \mathcal{A}|$ mutta $b(\mathcal{A}, k+1) < |\bigcup \mathcal{A}|$, saadaan samanlainen ristiriita.

Jos taas epäyhtälö 14 on tosi kaikilla k, parillisilla k pätee $b(\mathcal{A}, k) - |\bigcup \mathcal{A}| \leq b(\mathcal{A}, k) - b(\mathcal{A}, k+1) = S_{k+1}$ ja parittomilla k pätee $|\bigcup \mathcal{A}| - b(\mathcal{A}, k) \leq b(\mathcal{A}, k+1) - b(\mathcal{A}, k) = S_{k+1}$. Kummassakin tapauksessa epäyhtälö 13 on tosi.

Todistus. Inkluusio-ekskluusio -periaatteen nojalla $|\bigcup \mathcal{A}| = b(\mathcal{A}, n)$, ja siis $d(\mathcal{A}, k) = |b(\mathcal{A}, n) - b(\mathcal{A}, k)|$. Edelleen lemmoja 8 ja 9 käyttämällä saadaan

$$\begin{aligned} b(\mathcal{A}, n) - b(\mathcal{A}, k) &= \sum_{x \in U} \left(\sum_{j=0}^{n} (-1)^{j+1} \binom{s_x}{j} - \sum_{j=0}^{k} (-1)^{j+1} \binom{s_x}{j} \right) \\ &= \sum_{x \in U} \left((1-1)^n - (-1)^k \binom{s_x - 1}{k} \right) = (-1)^{k+1} \sum_{x \in U} \binom{s_x - 1}{k}. \end{aligned}$$

Jos $s_x \leq k$, varmasti $\binom{s_x - 1}{k} = 0$. Jos taas $s_x \geq k+1$, niin $\binom{s_x - 1}{k} \leq \frac{s_x}{k+1} \binom{s_x - 1}{k} = \binom{s_x}{k+1}$, joten $d(\mathcal{A}, k) = \sum_{x \in U} \binom{s_x - 1}{k} \leq \sum_{x \in U} \binom{s_x}{k+1} = S_{k+1}$. □

Toinen todistus. Kirjoitetaan $b(\mathcal{A}, k)$ alkuperäiseen muotoon, jossa ei esiinny tasojoukkoja:

$$\begin{aligned} b(\mathcal{A}, k) &= \sum \left((-1)^{|\mathcal{B}|+1} |\bigcap \mathcal{B}| \;\middle|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A},\ |\mathcal{B}| \leq k \right) \\ &= \sum_{x \in U} \sum \left((-1)^{|\mathcal{B}|+1} \;\middle|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A}_x,\ |\mathcal{B}| \leq k \right). \end{aligned}$$

Kuten ensimmäisessä todistuksessa, virhe $d(\mathcal{A}, k)$ voidaan kirjoittaa käyttämällä hyväksi tietoa $|\bigcup \mathcal{A}| = b(\mathcal{A}, n)$ muotoon

$$d(\mathcal{A}, k) = \left| \sum_{x \in U} \sum \left((-1)^{|\mathcal{B}|+1} \;\middle|\; \emptyset \neq \mathcal{B} \subseteq \mathcal{A}_x,\ |\mathcal{B}| > k \right) \right|. \tag{15}$$

Todistettavan epäyhtälön oikea puoli on vastaavassa muodossa

$$S_{k+1} = \sum_{x \in U} \sum \left(1 \;\middle|\; \emptyset \neq \mathcal{B} \in \mathcal{A}_x^{(k+1)} \right) = \sum_{x \in U} \left| \mathcal{A}_x^{(k+1)} \right|. \tag{16}$$

Riittää todistaa, että kunkin alkion $x \in U$ kontribuutio summaan (15) on pienempi tai yhtä kuin summaan (16). Jos $s_x = 0$ eli $\mathcal{A}_x = \emptyset$, selvästi kontribuutio kumpaankin summaan on 0. Oletetaan, että $\mathcal{A}_x \neq \emptyset$.

Olkoon $\mathcal{A}_x$:n indeksijoukko I_x. Olkoon $\mathcal{E}$ joukon I_x parilliskokoisten osajoukkojen perhe sekä $\mathcal{O}$ sen paritonkokoisten osajoukkojen perhe. Valitaan jokin $a \in I_x$ ja määritellään bijektio $f : \mathcal{E} \to \mathcal{O}$ kaavalla (9). Oletetaan nyt, että k on pariton; jos k on parillinen, seuraava todistus pätee, kun vaihdetaan keskenään joukot $\mathcal{E}$ ja $\mathcal{O}$ sekä muutetaan funktio f funktioksi f^{-1}.

Todistus perustuu siihen ajatukseen, että jos kuvaan 2 (sivulla 24) piirretään vaakaviiva tasojen k ja $k+1$ väliin, se leikkaa enintään niin monta nuolta kuin tasolla $k+1$ on joukkoja. Lasketaan kaavan (15) sisempi summa seuraavalla tavalla: Käydään läpi kaikki

$\mathcal{E}$:n joukot J, joille $|J| > k$. Jos myös $|f(J)| > k$, lasketaan summaan mukaan $(-1)^{|J|+1} + (-1)^{|f(J)|+1}$, mutta jos $|f(J)| \le k$, lasketaan vain $(-1)^{|J|+1}$.

Koska f on bijektio $\mathcal{E} \to \mathcal{O}$ ja $\mathcal{E} \cup \mathcal{O}$ on $\mathcal{P}(I_x)$:n ositus, näin tullaan laskeneeksi summa $\sum\big((-1)^{|J|+1} \mid J \subseteq I_x,\ |J| > k\big)$. Toisaalta jonojen $\mathcal{B} \subseteq \mathcal{A}_x$ ja joukkojen $J \subseteq I_x$ välillä vallitsee yksinkertainen vastaavuus: $|\mathcal{B}| = |J|$, kun J on jonon $\mathcal{B}$ indeksijoukko, joten näin saadaan todellakin haluttu summa $\sum\big((-1)^{|\mathcal{B}|+1} \mid \mathcal{B} \subseteq \mathcal{A}_x,\ |\mathcal{B}| > k\big)$.

Niitä $J \subseteq I_x$, joille sekä $|J| > k$ että $|f(J)| > k$, vastaava termi tässä laskentatavassa on 0, koska $J \in \mathcal{E}$ ja $f(J) \in \mathcal{O}$. Kaikki muut termit vastaavat sellaisia joukkoja J, että $|J| > k$ mutta $|f(J)| \le k$. Kuvauksen f konstruktiosta seuraa heti, että $|f(J)| = |J| \pm 1$, joten näille J pätee $|J| = k + 1$. Jokaista tällaista J vastaa termi $(-1)^{|J|+1}$, ja näitä termejä on enintään $\left|I_x^{(k+1)}\right| = \left|\mathcal{A}_x^{(k+1)}\right|$. □

Esimerkki 7. Esimerkin 4 katkaistut inkluusio-ekskluusio -summat ovat 18, $18 - 16 = 2$, $18 - 16 + 8 = 10$, $18 - 16 + 8 - 2 = 8$ ja $18 - 16 + 8 - 2 + 0 = 8$. Nähdään, että jokainen katkaistu summa poikkeaa todellisesta arvosta 8 enintään seuraavan tason verran.

Seuraava lause yleistää Bonferronin epäyhtälöä samaan tapaan kuin lause 10 inkluusio-ekskluusio -periaatetta.

Lause 12. *Olkoon* $0 \le k \le n$. *Olkoon lisäksi* ℓ *sellainen luku, että* $k \le \ell \le n$. *Merkitään*

$$a_k(\mathcal{A}, \ell) = \sum_{j=k}^{\ell} (-1)^{j-k} \binom{j-1}{k-1} S_j$$

ja

$$b_k(\mathcal{A}, \ell) = \sum_{j=k}^{\ell} (-1)^{j-k} \binom{j}{k} S_j.$$

Silloin

$$|\#P_k - a_k(\mathcal{A}, \ell)| \le \binom{\ell}{k-1} S_{\ell+1} \tag{17}$$

ja

$$|\#Q_k - b_k(\mathcal{A}, \ell)| \le \binom{\ell+1}{k} S_{\ell+1}. \tag{18}$$

Lause 10 on tämän erikoistapaus: kun $\ell = n$, $S_{\ell+1} = 0$ ja summat $a_{k,\ell}$ ja $b_{k,\ell}$ ovat tarkkoja. Galambos kutsuu tämän lauseen tuloksia klassisiksi Bonferronin epäyhtälöiksi.

Todistus. Aloitetaan jälleen joukkoa Q_k koskevasta tuloksesta. Edetään kuten lauseen 10 todistuksessa (sivulla 26). Ensin saadaan

$$b_k(\mathcal{A}, \ell) = \sum_{x \in U} \binom{s_x}{k} \sum_{j=k}^{\ell} (-1)^{j-k} \binom{s_x - k}{j - k}. \tag{19}$$

Ainoa ero lauseen 10 todistukseen on siinä, että sisempi summa lasketaan ℓ:ään eikä n:ään asti.

Lasketaan edelleen kutakin $x \in U$ vastaava termi summassa (19). Jos $s_x < k$, $\binom{s_x}{k} = 0$. Jos $s_x = k$ eli $x \in Q_k$, termiksi saadaan 1 aivan samasta syystä kuin lauseen 10 tapauksessa. Siis niiden x, joille $s_x \leq k$, osuus summassa (19) on täsmälleen sama kuin summassa (12). Jos $s_x > k$, alkiota x vastaava termi summassa (12) oli 0, mutta tässä tapauksessa termiksi tulee lemman 9 nojalla

$$\binom{s_x}{k} \sum_{j=k}^{\ell} (-1)^{j-k} \binom{s_x - k}{j - k} = \binom{s_x}{k} (-1)^{\ell-k} \binom{s_x - k - 1}{\ell - k}.$$

Epäyhtälön (18) vasen puoli on tämän perusteella

$$\sum \left(\binom{s_x}{k} \binom{s_x - k - 1}{\ell - k} \;\middle|\; x \in U,\, s_x > k \right).$$

Kirjoitetaan saman epäyhtälön oikea puoli lemman 8 ja lauseen 1(f) avulla muotoon

$$\binom{\ell + 1}{k} S_{\ell+1} = \sum_{x \in U} \binom{\ell + 1}{k} \binom{s_x}{\ell + 1} = \sum_{x \in U} \binom{s_x}{k} \binom{s_x - k}{\ell + 1 - k}.$$

Kummallakin puolella niitä $x \in U$ vastaavat termit, joille $s_x \leq \ell$, ovat nollia, joten riittää todistaa

$$\sum \left(\binom{s_x}{k} \binom{s_x - k - 1}{\ell - k} \;\middle|\; x \in U,\, s_x > \ell \right) \leq \sum \left(\binom{s_x}{k} \binom{s_x - k}{\ell + 1 - k} \;\middle|\; x \in U,\, s_x > \ell \right).$$

Tämä seuraa siitä, että

$$\binom{s_x - k - 1}{\ell - k} = \frac{\ell + 1 - k}{s_x - k} \binom{s_x - k}{\ell + 1 - k} \leq \binom{s_x - k}{\ell + 1 - k},$$

kun $s_x > \ell$.

Edellä esitetty epäyhtälön (18) todistus on olennaisesti sama kuin lähteessä [GS96] esitetty, ja saman lähteen mukaan epäyhtälö (17) todistetaan samaan tapaan. Edellisen todistuksen muuntaminen ei kuitenkaan näytä aivan suoraviivaiselta. Summa $a_k(\mathcal{A}, \ell)$ saadaan kyllä helposti muotoon

$$a_k(\mathcal{A}, \ell) = \sum_{x \in U} \sum_{j=k}^{\ell} (-1)^{j-k} \binom{j - 1}{k - 1} \binom{s_x}{j},$$

mutta lauseesta 1(f) ei ole nyt apua. Voidaan toki kirjoittaa

$$\binom{j - 1}{k - 1} = \frac{k}{j} \binom{j}{k},$$

mistä saadaan muoto

$$a_k(\mathcal{A},\ell) = \sum_{x\in U} \binom{s_x}{k} \sum_{j=k}^{\ell} (-1)^{j-k} \frac{k}{j} \binom{s_x-k}{j-k},$$

mutta sisempi summa näyttää olennaisesti vaikeammalta kuin lemman 9 tulos. Myöskään sellainen summaaminen kuin lauseen 10 todistuksessa ei tunnu auttavan.

Seuraava todistus on eräänlainen raa'an voiman ratkaisu. Kirjoitetaan epäyhtälön (17) vasen puoli, arvion $a_k(\mathcal{A},\ell)$ virhe, summaksi

$$\sum_{j=\ell+1}^{n} (-1)^{j-k} \binom{j-1}{k-1} S_j.$$

Jälleen lauseella 8 ja summausjärjestyksen vaihdolla saadaan virheelle esitys

$$\sum_{x\in U} v(x), \quad v(x) = \sum_{j=\ell+1}^{n} (-1)^{j-k} \binom{j-1}{k-1} \binom{s_x}{j}.$$

Esitetään epäyhtälön oikea puoli samoin muodossa

$$\sum_{x\in U} w(x), \quad w(x) = \binom{\ell}{k-1} \binom{s_x}{\ell+1}.$$

Nyt riittää todistaa $|v(x)| \le w(x)$ kaikille $x \in U$. Jos $s_x \le \ell$, selvästi $v(x) = w(x) = 0$, joten edelleen voidaan olettaa, että $s_x > \ell$. Samoin tiedetään, että kaikille $x \in U$ pätee $s_x \le n$. Koska $\binom{s_x}{j} = 0$, kun $j > s_x$, summan $v(x)$ ylärajaksi voidaan vaihtaa s_x.

Siirrytään tarkastelemaan funktioita V ja W, jotka määritellään kaavoilla

$$\begin{aligned} V(s) &= \sum_{j=\ell+1}^{s} (-1)^{j-k} \binom{j-1}{k-1} \binom{s}{j}, \\ W(s) &= \binom{\ell}{k-1} \binom{s}{\ell+1}. \end{aligned}$$

Nyt riittää todistaa $|V(s)| \le W(s)$ kaikilla $s > \ell$, koska määritelmän perusteella $v(x) = V(s_x)$ ja $w(x) = W(s_x)$. Tapauksessa $s = \ell+1$ saadaan

$$V(\ell+1) = (-1)^{\ell-k+1} \binom{\ell}{k-1} \quad \text{ja} \quad W(\ell+1) = \binom{\ell}{k-1}.$$

Todistetaan loput tapaukset induktiolla.

Liitteen A lemmasta 29 (sivulla 58) saadaan summalle $V(s)$ palautuskaava

$$V(s+1) - V(s) = (-1)^{\ell+k+1} \binom{\ell}{k-1} \binom{s}{\ell} \frac{\ell-k+1}{s-k+1}.$$

Funktiolle W on helppo johtaa lauseen 1(d) avulla vastaava rekursio:

$$W(s+1) - W(s) = \binom{\ell}{k-1}\left[\binom{s+1}{\ell+1} - \binom{s}{\ell+1}\right] = \binom{\ell}{k-1}\binom{s}{\ell}.$$

Oletuksesta $s > \ell$ seuraa $s - k + 1 > \ell - k + 1$. Mutta tämä riittää induktioaskeleen todistamiseksi: jos $|V(s)| \leq W(s)$,

$$|V(s+1)| \leq |V(s)| + |V(s+1) - V(s)| < W(s) + (W(s+1) - W(s)) = W(s+1). \quad \square$$

4.4 Todennäköisyystulkinta

Kappaleessa 4.2 todettiin, että joukon U voi tulkita äärelliseksi todennäköisyysavaruudeksi, jolloin esitetyillä tuloksilla on probabilistinen merkitys. Itse asiassa ne pätevät mielivaltaisessa todennäköisyysavaruudessa. Tämä johtuu seuraavasta lauseesta.

Lause 13 (Rényin lause). *Olkoon $\langle \Omega, \mathcal{B}, \mathsf{P} \rangle$ todennäköisyysavaruus eli Ω joukko, $\mathcal{B}$ sen sigma-algebra ja $\mathsf{P} : \mathcal{B} \to [0,1]$ todennäköisyysmitta. Olkoot k ja n positiivisia kokonaislukuja, ja olkoot $f_i : \mathcal{B}^n \to \mathcal{B}$, $i \in [k]$, n-paikkaisia Boolen funktioita[2] $\mathcal{B}$:ssä ja olkoot c_i, $i \in [k]$, reaalilukuja. Oletetaan, että epäyhtälö*

$$\sum_{i \in [k]} c_i \mathsf{P}(f_i(A_0, A_1, \ldots, A_{n-1})) \geq 0 \tag{20}$$

pätee aina, kun kukin A_i on joko Ω tai $\emptyset$. Silloin (20) pätee millä tahansa joukoilla $A_i \in \mathcal{B}$.

Todistus. Merkitään $\chi(A)$:lla joukon $A \in \mathcal{B}$ indikaattorifunktiota, ts.

$$\chi(A)(\omega) = \begin{cases} 1, & \text{jos } \omega \in A, \\ 0 & \text{jos } \omega \notin A. \end{cases}$$

Edelleen merkitään $\chi_i = \chi(f_i(A_0, A_1, \ldots, A_{n-1}))$. Olkoot $A_j \in \mathcal{B}$, $j \in [n]$, mielivaltaisia joukkoja. Todistetaan epäyhtälö

$$\sum_{i \in [k]} c_i \chi_i(\omega) \geq 0$$

kaikilla $\omega \in \Omega$. Väite seuraa tästä odotusarvo-operaattorin E ominaisuuksien avulla. Ensinnäkin $\mathsf{E}\left(\sum_{i \in [k]} c_i \chi_i\right) \geq 0$, ja odotusarvon lineaarisuuden nojalla $\sum_{i \in [k]} c_i \mathsf{E}(\chi_i) \geq 0$. Toiseksi kaikille joukoille $A \in \mathcal{B}$ pätee $\mathsf{P}(A) = \mathsf{E}(\chi(A))$, joten $\mathsf{E}(\chi_i) = \mathsf{P}(f_i(A_0, A_1, \ldots, A_{n-1}))$.

Olkoon siis $\omega \in \Omega$. Asetetaan kaikilla $j \in [n]$

$$A_j^* = \begin{cases} \emptyset, & \text{jos } \omega \notin A_j \\ \Omega, & \text{jos } \omega \in A_j. \end{cases}$$

[2]Lyhyesti sanottuna: operaatioista $\cup$, $\cap$ ja $\complement$ (komplementti) rekursiivisesti koottuja funktioita.

Lauseen oletuksen nojalla epäyhtälö

$$\sum_{i\in[k]} c_i \mathsf{P}(f_i(A_0^*,\dots,A_{n-1}^*)) \geq 0$$

on tosi. Joukkojen A_j^* valinnasta seuraa toisaalta, että jokainen joukko $f_i(A_0^*,\dots,A_{n-1}^*)$ on joko $\emptyset$ tai Ω (tämän voi todistaa yksinkertaisella induktiolla Boolen funktion f_i rakenteen suhteen). Siis

$$\mathsf{P}(f_i(A_0^*,\dots,A_{n-1}^*)) = \chi(f_i(A_0^*,\dots,A_{n-1}^*))(\omega) = \chi_i(\omega).$$

Viimeinen yhtäsuuruus johtuu siitä, että konstruktion perusteella alkio ω kuuluu joukkoon A_j^* jos ja vain jos se kuuluu joukkoon A_j. □

Huomautus 5. Lause 13 ja sen todistus pysyvät voimassa, jos vaihdetaan jokaisen epäyhtälön $X \geq 0$ tilalle yhtälö $X = 0$.

Esimerkiksi inkluusio-ekskluusio -periaate ja Bonferronin epäyhtälö voitaisiin todistaa triviaaliavaruudessa $\langle \Omega, \{\emptyset, \Omega\}, \mathsf{P}\rangle$. Sitten Rényin lauseesta seuraa heti, että tulokset pätevät kaikissa todennäköisyysavaruuksissa. Äärellisen avaruuden $\langle U, \mathcal{P}(U), \mathsf{P}\rangle$, $\mathsf{P}(X) = \#X/\#U$, tapauksessa tämä menetelmä vastaa sitä, että vaihdetaan summausjärjestystä ja todistetaan epäyhtälö erikseen jokaiselle $x \in U$. Ajatus on samantapainen kuin lauseen 6 (sivulla 15) todistuksen alku.

Lähteessä [GS96] nimitetään *Bonferroni-tyyppisiksi* epäyhtälöiksi sellaisia kaikissa todennäköisyysavaruuksissa päteviä lineaarisia epäyhtälöitä, jotka koskevat äärellistä tapahtumajonoa $\mathcal{A}$ ja joiden muuttujat ovat seuraavien tapahtumien todennäköisyyksiä:

- leikkaukset $\bigcap \mathcal{B}$, missä $\mathcal{B} \subseteq \mathcal{A}$ (vrt. summan S_j termit),
- "vähintään k $\mathcal{A}$:n tapahtumaa toteutuu" (vrt. P_k),
- "tasan k $\mathcal{A}$:n tapahtumaa toteutuu" (vrt. Q_k).

Kaikki nämä tapahtumat ovat sellaista muotoa, että lausetta 13 voidaan soveltaa.

5 Verkkoihin perustuvia epäyhtälöitä

Edellä esitellyt tulokset eivät auta luvussa 3 esiteltyyn ongelmaan: siinä tunnetaan joukkojen A_i leikkauksia eri tavalla kuin lauseen 11 tarvitsemina symmetrisinä summina S_j. Tässä luvussa esitellään menetelmä, joka käyttää niiden sijasta tai lisäksi eräänlaisia epäsymmetrisiä summia.

Epäsymmetristen summien käyttämistä voi perustella silloinkin, kun tunnetaan täsmälleen sellaiset leikkaukset, että voidaan laskea summat S_j. Lauseesta 11 saadaan ylä- tai alaraja sen mukaan, onko käytössä parillinen vai pariton määrä summia. Jos halutaan molemmat, jompaakumpaa laskettaessa jätetään seuraavan tason tiedot käyttämättä. Esimerkiksi siinä yksinkertaisessa tapauksessa, että tunnetaan vain joukkojen A_i ja niiden parittaisten leikkausten $A_i \cap A_j$ koot, saadaan epäyhtälöt

$$\sum |A_i| - \sum_{i<j} |A_i \cap A_j| \leq |\bigcup A_i| \leq \sum |A_i| \,. \tag{21}$$

Oikeanpuoleinen epäyhtälö on suorastaan triviaali. Sen oikealta puolelta voi tämän luvun keinoilla vähentää useita leikkauksia $A_i \cap A_j$.

Seuraavassa esitettävän menetelmän nimi on englanniksi *graph sieve*, jonka sananmukainen suomennos olisi "verkkoseula". Sana "seula" tulee lukuteoriasta, jossa inkluusio-ekskluusio -periaate nähdään Eratostheneen seulan yleistyksenä. Epäyhtälön termit määräytyvät sopivasti valitusta verkosta tai hyperverkosta.

Rényin lause 13 pätee myös tämän luvun tuloksiin. Koska käytetään epäsymmetrisiä summia, tulokset voivat olla erilaisia jonon $\mathcal{A}$ eri järjestyksillä. Näin jokaisesta epäyhtälöstä saadaan useita versioita, joista jokainen on kyllä voimassa perusjoukon U ja jonon $\mathcal{A}$ valinnoista riippumatta, mutta joista voidaan valita se, joka on tiukin käsillä olevassa tapauksessa.

Kappale 5.2 perustuu enimmäkseen artikkeliin [Wor82], joskin vastaavia tuloksia on kaiketi esitetty aiemminkin. Kappaleen 5.3 menetelmä on peräisin kirjasta [GS96]. Kappaleen 5.4 hyperverkkoihin liittyvät epäyhtälöt ovat artikkeleista [Gra93] ja [Gra94].

5.1 Verkot

Seuraavassa esitellään verkot ja eräitä niihin liittyviä käsitteitä. Lukijan olisi syytä tuntea jonkin verran verkkoteoriaa, sillä seuraava käsittely on pelkästään tiivis luettelo käytetyistä määritelmistä ja tuloksista, eikä se sisällä todistuksia.

Määritelmä 14. Olkoon V epätyhjä äärellinen joukko ja $E \subseteq V^{(2)}$ joukko V:n kaksialkioisia osajoukkoja eli järjestämättömiä pareja. Silloin paria $G = \langle V, E \rangle$ sanotaan *verkoksi*, V:tä verkon G *solmujen joukoksi* ja E:tä G:n *kaarien joukoksi*. Jos $G = \langle X, Y \rangle$ on verkko, merkitään $V(G) = X$ ja $E(G) = Y$.

Huomautus 6. Määritelmän 14 mukaan verkot ovat suuntaamattomia, eikä solmusta v voi olla kaarta itseensä.

Määritelmä 15. Olkoon G verkko, $\emptyset \neq V' \subseteq V(G)$ ja $E' \subseteq E(G) \cap (V')^{(2)}$. Silloin verkko $\langle V', E' \rangle$ on verkon G *aliverkko*. Jos edelleen $E' = E(G) \cap (V')^{(2)}$, verkko $\langle V', E' \rangle$ on *joukon V' indusoima G:n aliverkko*.

Määritelmä 16. Olkoon G verkko. Jos $v_0, v_1, \ldots, v_{s-1}$ ovat $V(G)$:n alkioita ja $\{v_i, v_{i+1}\} \in E(G)$ kaikilla $i \in [s-1]$, sanotaan jonoa $(v_0, v_1, \ldots, v_{s-1})$ verkon G *poluksi*.

Määritelmä 17. Olkoon G verkko. Jos kaikille eri solmuille $v, v' \in V(G)$, $v \neq v'$, on olemassa polku $(v_0, \ldots, v_{s-1})$, missä $v_0 = v$ ja $v_{s-1} = v'$, verkkoa G sanotaan *yhtenäiseksi*.

Määritelmä 18. Olkoon G verkko ja $\emptyset \neq V' \subseteq V(G)$. Jos G:ssä ei ole kaaria $\{v, v'\}$, joille $v' \in V'$ ja $v \notin V'$, ja joukon V' indusoima G:n aliverkko on yhtenäinen, tätä aliverkkoa sanotaan verkon G *komponentiksi*. Jos $V' \subsetneq V(G)$, komponenttia sanotaan *aidoksi*.

Lemma 14. *Olkoon G yhtenäinen verkko. Silloin G:n ainoa komponentti on G itse.*

Määritelmä 19. Olkoon G verkko ja olkoon $(v_0, v_1, \ldots, v_{s-1})$ G:n polku. Jos $\{v_{s-1}, v_0\} \in E(G)$, sanotaan, että $(v_0, v_1, \ldots, v_{s-1})$ on verkon G *silmukka*. Jos verkossa G ei ole yhtään silmukkaa, sanotaan G:tä *silmukattomaksi*.

Määritelmä 20. Olkoon G yhtenäinen silmukaton verkko. Silloin G:tä sanotaan *puuksi*.

Lemma 15. *Olkoon G puu, jossa on s solmua, $|V(G)| = s$. Silloin G:ssä on $s-1$ kaarta, $|E(G)| = s - 1$.*

5.2 Yksinkertainen tapaus

Artikkeli [Wor82] tarkastelee epäyhtälön (21) parantamista seuraavasti. Muodostetaan verkko $G = \langle [n], E \rangle$, jonka solmuja ovat $\mathcal{A}$:n joukkojen indeksit ja jonka kaaret valitaan seuraavalla säännöllä: $\{i, j\} \in E$ jos ja vain jos $A_i \cap A_j \neq \emptyset$. Jos H on G:n silmukaton aliverkko, epäyhtälön (21) oikeaa puolta voidaan parantaa vähentämällä H:n kaaria $\{i, j\}$ vastaavat parittaiset leikkaukset $A_i \cap A_j$.

Tuloksen voi perustella intuitiivisesti seuraavasti: Jos alkio $x \in U$ kuuluu H:n kaarta $\{i, j\}$ vastaavaan leikkaukseen $A_i \cap A_j$, se lasketaan alkuperäiseen summaan kaksi kertaa, joten sen voi vähentää kerran. Jos alkio kuuluu useampaan kuin yhteen H:n kaareen, se on lisätty alkuperäiseen summaan vielä useampia kertoja. Vaatimus H:n silmukattomuudesta takaa, ettei alkiota vähennetä liian monta kertaa.

Seuraavassa esitetään ja todistetaan tulos täsmällisemmin.

Lause 16. *Olkoon $G = \langle [n], E \rangle$ verkko, jossa solmujen i ja j välillä on kaari $\{i, j\}$, jos ja vain jos $A_i \cap A_j \neq \emptyset$. Olkoon H silmukaton G:n aliverkko. Seuraava epäyhtälö pätee:*

$$\left|\bigcup \mathcal{A}\right| \leq \sum_{A \in \mathcal{A}} |A| - \sum_{\{i,j\} \in E(H)} |A_i \cap A_j| . \tag{22}$$

Todistus. Koska H on silmukaton, sen komponentit ovat puita. Jos näitä komponentteja on enemmän kuin yksi, $E(H)$:hon voidaan lisätä ylimääräinen kaari niin, että H pysyy silmukattomana ja sen komponenttien määrä vähenee yhdellä. Tällöin epäyhtälön (22) oikea puoli voi vain pienentyä tai pysyä samana. Kun tätä operaatiota toistetaan, H:sta tulee puu. Riittää siis todistaa väite puille H. Edelleen voidaan olettaa, että H:n solmujen joukko on $[n]$; muussa tapauksessa voidaan jälleen pienentää epäyhtälön (22) oikeaa puolta lisäämällä H:hon solmuja ja sopivia kaaria.

Permutoidaan joukkojen $A_j \in \mathcal{A}$ indeksit seuraavasti: Asetetaan $p_0 = 0$. Kun p_{j-1} on valittu, $j < n$, valitaan p_j sellaiseksi, että $\{p_i, p_j\} \in E(H)$ jollain $i < j$, ja merkitään $j' = i$. Jos tämä ei ole mahdollista, joukon $\{p_0, p_1, \ldots, p_{j-1}\}$ indusoima aliverkko on H:n aito komponentti, joten H ei ole yhtenäinen eikä siis puu.

Nyt voidaan kirjoittaa

$$\bigcup_{j\in[n]} A_{p_j} = A_{p_0} \cup \bigcup_{0\neq j\in[n]} \big(A_{p_j} \setminus A_{p_{j'}}\big).$$

Sovelletaan oikeanpuoleiseen yhdisteeseen Bonferronin epäyhtälöä (lause 11) k:n arvolla 0. Yhdisteen koolle saadaan sen avulla yläraja

$$\begin{aligned}\big|A_{p_0}\big| + \sum_{0\neq j\in[n]} \big|A_{p_j} \setminus A_{p_{j'}}\big| &= \big|A_0\big| + \sum_{0\neq j\in[n]} \Big(\big|A_{p_j}\big| - \big|A_{p_j} \cap A_{p_{j'}}\big|\Big)\\ &= \sum_{j\in[n]} \big|A_{p_j}\big| - \sum_{0\neq j\in[n]} \big|A_{p_j} \cap A_{p_{j'}}\big|.\end{aligned}$$

Viimeisen summan jokaista $n-1$ termiä vastaa jokin H:n kaari $\{j', j\}$, ja toisaalta summan eri termejä vastaavat eri kaaret. Mutta koska H on puu, siinä on tasan $n-1$ kaarta, joten summassa esiintyvät kaikki H:n kaaret. On todistettu epäyhtälö (22). □

Huomautus 7. Kuten tämän luvun alussa todettiin, epäsymmetrisiä summia käytettäessä voidaan valita $\mathcal{A}$:lle kulloinkin paras järjestys. Tässä tapauksessa se voidaan löytää antamalla G:n kaarelle $\{i, j\}$ paino $|A_i \cap A_j|$ ja etsimällä ns. Kruskalin algoritmilla kullekin G:n komponentille maksimaalinen virittävä puu.

Esimerkki 8. Tarkastellaan esimerkkiä 4 (sivulla 22). Siihen liittyvästä taulukosta 3 nähdään, että verkossa G ovat kaikki muut mahdolliset särmät paitsi $\{1, 2\}$. Jos nyt valitaan verkkoon H särmät $\{0, 1\}$, $\{0, 3\}$, $\{0, 4\}$ ja $\{2, 3\}$, saadaan tulos

$$|\textstyle\bigcup \mathcal{A}| \leq 18 - 1 - 3 - 3 - 2 = 9.$$

Todistuksesta nähdään myös helposti, että aliverkossa H voisi olla myös sellaisia kaaria, jotka puuttuvat G:stä: ne tuottavat tietysti vain nollia vähennettävään summaan. Verkko H voidaan siis valita myös riippumatta siitä, millä joukoista A_j on epätyhjä leikkaus. Eräs mahdollinen puu on *tähti* H_j, $j \in [n]$: $E(H_j) = \{\, \{i, j\} \mid i \in [n], i \neq j\,\}$. Se tuottaa epäyhtälön

$$|\textstyle\bigcup \mathcal{A}| \leq \displaystyle\sum_{i\in[n]} |A_i| - \sum\Big(|A_i \cap A_j| \;\Big|\; i \in [n],\, i \neq j\Big). \tag{23}$$

Tästä epäyhtälöstä saadaan myös symmetrisiä summia koskeva tulos:

Lause 17. *Seuraava epäyhtälö pätee.*

$$|\bigcup \mathcal{A}| \leq S_1 - \frac{2}{n} S_2.$$

Todistus. Kirjoitetaan epäyhtälö (23) kaikille n:lle tähdelle H_j ja lasketaan keskiarvo. Koska $2S_2 = \sum\bigl(\#A_i \cap A_j \bigm| i, j \in [n], i \neq j\bigr)$, saadaan väite. □

5.3 Yleistys

Edellisen päättelyn yleistämiseksi tutkitaan summaa (22) jälleen alkioittain. Määritellään jokaiselle $x \in \bigcup \mathcal{A}$ solmujoukko V_x säännöllä $j \in V_x \iff x \in A_j$. Olkoon $H_x = \langle V_x, E_x \rangle$ joukon V_x indusoima aliverkko. Tätä x kohti summataan epäyhtälön (22) oikealla puolella $|V_x| - |E_x|$. Toisin sanoen

$$\sum_{i \in [n]} |A_i| - \sum\Bigl(|A_i \cap A_j| \Bigm| \{i, j\} \in E(H),\, i \neq j\Bigr) = \sum_{x \in \bigcup \mathcal{A}} \bigl(|V_x| - |E_x|\bigr).$$

Koska H on silmukaton, sen aliverkolle H_x pätee, että $|E_x| \leq |V_x| - 1$, joten alkiota x vastaava termi on yhteensä vähintään 1. Siis edellinen summa on suurempi tai yhtä suuri kuin

$$\sum_{x \in \bigcup \mathcal{A}} 1 = |\bigcup \mathcal{A}|\,.$$

Tästä saadaan siinä mielessä suoraviivaisempi todistus lauseelle 16, että verkon H silmukattomuutta käytetään hyväksi luonnollisella tavalla ilman järjestämisajatusta. Ajattelutavasta on hyötyä, kun halutaan muuntaa lausetta hieman.

Jos nimittäin lisätään verkkoon H ylimääräinen kaari $\{k, \ell\}$, siihen saattaa tulla silmukka, mikä tietysti kaataa lauseelle 16 ensin esitetyn todistuksen. Alkioittaiseen todistukseen voi tehdä pienen muutoksen: Ensinnäkin jokaista $x \in \bigcup \mathcal{A}$ vastaava termi on vähintään 0. Jos x toisaalta kuuluu tasan yhteen joukkoon $A \in \mathcal{A}$, välttämättä $|E_x| = 0$, joten x:n kontribuutio summaan on tasan 1. Saadaan siis epäyhtälö, joka koskee joukkoa Q_1 (vrt. lause 10).

Lause 18. *Olkoot G ja H kuten lauseessa 16, ja olkoot $k, \ell \in [n]$ sellaiset alkiot, että $\{k, \ell\} \notin E(H)$. Silloin*

$$|Q_1| \leq \sum_{A \in \mathcal{A}} |A| \;-\; \sum_{\{i,j\} \in E(H)} |A_i \cap A_j| \;-\; |A_k \cap A_\ell|\,.$$

Kokeillaan epäyhtälön (23) todistamista suoraan alkioittain. Käytetään tähden H_j sijasta komplementtiverkkoa $H_j^* = \langle V_j, E_j^* \rangle$, jossa mistään alkiosta $i \neq j$ ei ole kaarta j:hin mutta kaikki muut mahdolliset kaaret ovat mukana. Tämän verkon aliverkoilla $\langle V', E' \rangle$ on seuraava ominaisuus: $|V'| - \bigl|(V')^{(2)} \setminus E'\bigr| \geq 1$. Toisin sanoen aliverkoissa on aidosti enemmän solmuja kuin sellaisia solmupareja, joita ei yhdistä kaari. Alkioittaisessa tarkastelussa tämä merkitsee, että joihinkin joukkoihin A_i kuuluva x tuottaa epäyhtälön (23) oikealle puolelle vähintään ykkösen, ja kun summataan kaikkien sellaisten x suhteen, saadaan juuri tämä epäyhtälö.

Käyttämällä tähän tapaan verkkojen ominaisuuksia saadaan menetelmä epäyhtälöiden todistamiseksi. Yleisyyden nimissä menetelmässä on paljon vapaita parametreja, joten ymmärtämisen helpottamiseksi voi olla hyvä verrata esitystä edellä esitettyyn lauseen 16 alkioittaiseen todistukseen. Seuraavassa käytetään esimerkkinä sellaisia parametrien arvoja, että menetelmä tuottaa edellisen todistuksen.

Olkoon $G = \langle [n], E \rangle$ verkko. Olkoon $(\, \ell_j \mid j \in [m] \,)$ jono luonnollisia lukuja, joka määrää sen, kuinka monen joukon leikkausten kokoja summataan (esimerkissä $m = 2$ ja $\ell_0 = 1, \ell_1 = 2$). Olkoon edelleen $(\, p_j \mid j \in [m])$ jono verkkojen solmujoukkojen ominaisuuksia: kukin p_j on kaksipaikkainen totuusarvoinen funktio, jonka ensimmäinen argumentti on verkko ja toinen argumentti joukko saman verkon solmuja. (Esimerkissä p_0 on identtisesti tosi, ja p_1 on tosi, jos ja vain jos toinen argumentti on verkon särmä.) Kun $h \subseteq [n]$, merkitään $\mathcal{H}_j(h)$:lla niiden h:n osajoukkojen perhettä, joissa on ℓ_j alkiota ja joille p_j on tosi:

$$\mathcal{H}_j(h) = \{\, h' \in h^{(\ell_j)} \mid p_j(G, h') \,\}.$$

(Esimerkissä siis $\mathcal{H}_0(h)$ on h:n yksiöiden joukko ja $\mathcal{H}_1(h)$ on h:n indusoiman G:n aliverkon kaarien joukko.) Nyt voidaan määritellä tason j epäsymmetrinen summa

$$S_{(j)} = \sum \Big(|\bigcap_{i \in J} A_i| \;\Big|\; J \in \mathcal{H}_j([n]) \Big).$$

(Esimerkissä $S_{(0)} = \sum_{A \in h} |A|$ ja $S_{(1)} = \sum \big(|A \cap B| \;\big|\; A, B \in h, (A, B) \in E \big)$, missä $h = [n]$.) Olkoon vielä $(\, a_j \mid j \in [m] \,)$ jono reaalilukuja (esimerkissä $a_0 = 1$, $a_1 = -1$). Tavoitteena on todistaa epäyhtälö

$$\sum_{j=0}^{m-1} a_j S_{(j)} \geq |P_k| \,, \tag{24}$$

missä luku $k \in \mathbb{Z}^+$ määrää osajoukkohilan sen tason, jonka yhdisteelle arviota ollaan tekemässä (esimerkissä $k = 1$).

Lause 19. *Jos epäyhtälö*

$$\sum_{j=0}^{m-1} a_j \, |\mathcal{H}_j(h)| \geq \begin{cases} 1, & |h| \geq k, \\ 0, & |h| < k \end{cases} \tag{25}$$

pätee kaikilla $h \subseteq [n]$, niin myös epäyhtälö (24) pätee.

Todistus. Kirjoitetaan $S_{(j)}$ summana alkioiden $x \in U$ suhteen. Olkoon sitä varten I_x niiden $i \in [n]$ joukko, joille $x \in A_i$.

$$\begin{aligned} S_{(j)} &= \sum\Big(\big|\bigcap_{i \in J} A_i\big| \;\Big|\; J \in \mathcal{H}_j([n])\Big) \\ &= \sum\Big(\sum(1 \mid x \in \bigcap_{i \in J} A_i) \;\Big|\; J \in \mathcal{H}_j([n])\Big) \\ &= \sum_{x \in U} |\mathcal{H}_j([n]) \cap \mathcal{P}(I_x)| = \sum_{x \in U} |\mathcal{H}_j(I_x)|\,. \end{aligned}$$

Joukossa P_k ovat täsmälleen ne $x \in U$, joille $s_x = |I_x| \geq k$, joten väitteen 24 oikea puoli on summa

$$|P_k| = \sum_{x \in U} \chi(|I_x| \geq k),$$

missä $\chi(|I_x| \geq k) = 1$, jos $|I_x| \geq k$, muuten 0. Nyt voidaan kirjoittaa väite muotoon

$$\sum_{x \in U} \sum_{j=0}^{m-1} a_j\, |\mathcal{H}_j(I_x)| \geq \sum_{x \in U} \chi(|I_x| \geq k).$$

Väite seuraa epäyhtälöstä (25), joka pätee oletuksen mukaan kaikilla $h \subseteq [n]$. $\square$

Esimerkissä siis

$$\begin{aligned} &\ell_0 = 1, \qquad \mathcal{H}_0(h) = \{\, \{i\} \mid i \in h \,\}, && a_0 = 1, \\ &\ell_1 = 2, \qquad \mathcal{H}_1(h) = \{\, \{r,s\} \mid \{r,s\} \in h^{(2)} \cap E(H) \,\}, && a_1 = -1. \end{aligned}$$

Ehdon 25 vasen puoli on tässä tapauksessa

$$|\mathcal{H}_0(h)| - |\mathcal{H}_1(h)| = |h| - \left|E(H) \cap h^{(2)}\right|.$$

Ehto toteutuu, sillä epätyhjän joukon h indusoimassa silmukattoman verkon aliverkossa on solmuja vähintään yksi enemmän kuin kaaria.

Lause 20. *Jos lausetta 19 muutetaan seuraavasti, se pätee edelleen: vaihdetaan P_k:n tilalle Q_k ja epäyhtälön 25 oikeaksi puoleksi*

$$\begin{cases} 1, & |h| = k, \\ 0, & |h| \neq k. \end{cases}$$

Todistus. Joukon Q_k koon kertoo yhtälö

$$|Q_k| = \sum_{x \in U} \chi(|I_x| = k),$$

missä $\chi(|I_x| = k) = 1$, jos $|I_x| = k$, muuten 0. Muilta osin todistus etenee aivan samalla tavalla. $\square$

Huomautus 8. Lauseet 19 ja 20 pätevät edelleen, jos niitä muutetaan vaihtamalla epäyhtälöiden (24) ja (25) merkeiksi $\leq$.

5.4 Hyperverkot

Jos edellisen yleistyksen tavalla halutaan todistaa monimutkaisempia kuin kaksitasoisia epäyhtälöitä, voidaan käyttää verkon G sellaisia ominaisuuksia, jotka liittyvät useampaan kuin kahteen solmuun. Toinen, ehkä luonnollisempi, tapa edetä lauseesta 16 on tarkastella verkkojen sijasta hyperverkkoja.

Määritelmä 21. Kun $\alpha \in \mathbb{Z}^+$, *α-uniforminen hyperverkko* on pari $H = \langle V, E \rangle$, missä V on äärellinen joukko ja $E \subseteq V^{(\alpha)}$ on perhe V:n α-alkioisia osajoukkoja.

Voitaisiin määritellä myös α-hyperpuu yleistämään puita. Puiden tärkeä ominaisuus on edellisen kappaleen perusteella särmien vähyys siinä mielessä, että jokaisessa epätriviaalissa aliverkossa on enemmän solmuja kuin särmiä, ja osoittautuu, että tämän piirteen mielekäs yleistys johtaa laajempaan verkkojen luokkaan kuin eräässä mielessä luonnolliset α-hyperpuut (esim. [Gra93]).

Määritelmä 22. Olkoon $H = \langle V, E \rangle$ α-uniforminen hyperverkko. Sanotaan H:ta *harvaksi*, jos jokaisen osajoukon $I \subseteq V$ indusoima alihyperverkko $\langle I, E_I \rangle = \langle I, E \cap I^{(\alpha)} \rangle$ toteuttaa ehdon $|E_I| \leq \binom{|I|-1}{\alpha-1}$.

Kuten huomautuksessa 4 (sivulla 28) todettiin, Bonferronin epäyhtälön voi muotoilla niin, että inkluusio-ekskluusio -osasummat ovat vuorotellen luvun $|\bigcup \mathcal{A}|$ ylä- ja alarajoja. Seuraava lause tiukentaa näitä rajoja termillä, joka lasketaan harvan hyperverkon H avulla.

Määritelmä 23. Olkoon $H = \langle V, E \rangle$ α-uniforminen hyperverkko. Silloin

$$S_{(H)} = \sum \Big(\big| \bigcap_{j \in J} A_j \big| \ \Big| \ J \in E \Big).$$

Lause 21. *Olkoon $0 \leq \alpha \leq n$, ja olkoon $H = \langle [n], V \rangle$ α-uniforminen harva hyperverkko. Jos α on pariton,*

$$|\bigcup \mathcal{A}| \geq \sum_{j=1}^{\alpha-1} (-1)^{j+1} S_j + S_{(H)},$$

ja jos α on parillinen,

$$|\bigcup \mathcal{A}| \leq \sum_{j=1}^{\alpha-1} (-1)^{j+1} S_j - S_{(H)}.$$

Huomautus 9. Väitteen voi ilmaista lyhyemmin (mutta ehkä epäselvemmin) muodossa

$$(-1)^{\alpha+1} |\bigcup \mathcal{A}| \geq (-1)^{\alpha+1} \sum_{j=1}^{\alpha-1} (-1)^{j+1} S_j + S_{(H)}. \tag{26}$$

Todistus. Todistetaan huomautuksen 9 muotoilu (26). Merkitään kaikilla $J \subseteq [n]$, $J \neq \emptyset$,

$$B_J = \Big(\bigcap_{j\in J} A_j\Big) \setminus \Big(\bigcup_{j\in[n]\setminus J} A_j\Big).$$

Saadaan joukon $\bigcup \mathcal{A}$ ositus: $\bigcup \mathcal{A} = \bigcup\big(B_J \mid \emptyset \neq J \subseteq [n]\big)$ ja $B_J \cap B_{J'} = \emptyset$ kaikilla $J \neq J'$. Edelleen $\bigcap_{j\in J} A_j = \bigcup\big(B_I \mid J \subseteq I \subseteq [n]\big)$ kaikilla $\emptyset \neq J \subseteq [n]$.

Väitteen hyperverkkotermille saadaan summausjärjestystä vaihtamalla

$$\begin{aligned}
S_{(H)} &= \sum\Big(|\bigcap_{j\in J} A_j| \;\Big|\; J \in E\Big)\\
&= \sum\Big(\sum\big(\#B_I \mid J \subseteq I \subseteq [n]\big) \;\Big|\; J \in E\Big)\\
&= \sum\Big(\sum\big(\#B_I \mid J \in E,\, J \subseteq I)\big) \;\Big|\; \emptyset \neq I \subseteq [n]\Big)\\
&\leq \sum\left(|B_I| \binom{|I|-1}{\alpha-1} \;\middle|\; \emptyset \neq I \subseteq [n]\right).
\end{aligned}$$

Epäyhtälö seuraa siitä, että H on harva.

Inkluusio-ekskluusio -termin käsittelemiseksi kirjoitetaan

$$\begin{aligned}
S_j &= \sum\Big(|\bigcap_{i\in J} A_i| \;\Big|\; J \in [n]^{(j)}\Big)\\
&= \sum\Big(\sum\big(|B_I| \mid J \subseteq I \subseteq [n]\big) \;\Big|\; J \in [n]^{(j)}\Big)\\
&= \sum\Big(|B_I| \sum\big(1 \mid J \subseteq I,\, |J| = j\big) \;\Big|\; \emptyset \neq I \subseteq [n]\Big)\\
&= \sum\left(|B_I| \binom{|I|}{j} \;\middle|\; \emptyset \neq I \subseteq [n]\right).
\end{aligned}$$

Lemman 9 (sivulla 25) avulla saadaan

$$\sum_{j=1}^{\alpha-1} (-1)^{j+1} \binom{|I|}{j} = -\sum_{j=0}^{\alpha-1} (-1)^j \binom{|I|}{j} + \binom{|I|}{0} = (-1)^\alpha \binom{|I|-1}{\alpha-1} + 1.$$

Siis inkluusio-ekskluusio -termi on

$$\begin{aligned}
(-1)^{\alpha+1} \sum_{j=1}^{\alpha-1} (-1)^{j+1} S_j &= (-1)^{\alpha+1} \sum\left(|B_I| \sum_{j=1}^{\alpha-1} (-1)^{j+1} \binom{|I|}{j} \;\middle|\; \emptyset \neq I \subseteq [n]\right)\\
&= (-1)^{\alpha+1} \sum\left(|B_I| \left[(-1)^\alpha \binom{|I|-1}{\alpha-1} + 1\right] \;\middle|\; \emptyset \neq I \subseteq [n]\right)\\
&= \sum\left(|B_I| \left[-\binom{|I|-1}{\alpha-1} + (-1)^{\alpha+1}\right] \;\middle|\; \emptyset \neq I \subseteq [n]\right).
\end{aligned}$$

Epäyhtälön (26) oikeaa puolta voidaan nyt arvioida ylöspäin:

$$
\begin{aligned}
(-1)^{\alpha+1}&\sum_{j=1}^{\alpha-1}(-1)^{j+1}S_j + S_{(H)} \\
&\leq \sum\left(|B_I|\left[-\binom{|I|-1}{\alpha-1} + (-1)^{\alpha+1} + \binom{|I|-1}{\alpha-1}\right] \;\middle|\; \emptyset \neq I \subseteq [n]\right) \\
&\qquad\qquad = (-1)^{\alpha+1}\sum\left(|B_I| \;\middle|\; \emptyset \neq I \subseteq [n]\right) = (-1)^{\alpha+1}\left|\bigcup \mathcal{A}\right|. \qquad \square
\end{aligned}
$$

Samaan tapaan kuin aiemmin, tulos voidaan yleistää koskemaan joukkoja P_k ja Q_k. Joukkoa Q_k koskeva tulos on tässäkin tapauksessa yksinkertaisempi.

Lause 22. *Olkoot $k, \alpha \in \mathbb{Z}^+$ luvut, joille $k < \alpha$. Olkoon $H = \langle [n], E\rangle$ α-uniforminen hyperverkko, joka toteuttaa seuraavan ominaisuuden: minkä tahansa joukon $I \subseteq [n]$ indusoimassa alihyperverkossa on enintään*

$$\binom{|I|}{k}\binom{|I|-k-1}{\alpha-k-1}$$

kaarta. Silloin

$$(-1)^{\alpha+k}|Q_k| \geq (-1)^{\alpha+k}\sum_{j=k}^{\alpha-1}(-1)^{j+k}\binom{j}{k}S_j + S_{(H)}. \tag{27}$$

Todistus. Kuten lauseen 10 todistuksessa (sivulla 26), kirjoitetaan

$$(-1)^{\alpha+k}\sum_{j=k}^{\alpha-1}(-1)^{j+k}\binom{j}{k}S_j = \sum_{x\in U}\binom{s_x}{k}\sum_{j=k}^{\alpha-1}(-1)^{j+\alpha}\binom{s_x-k}{j-k}. \tag{28}$$

Tarkastellaan mielivaltaista $x \in U$ vastaavaa termiä summassa (28). Jos $s_x < k$, termi on 0. Tapauksessa $s_x = k$ termi on $(-1)^{\alpha+k}\binom{s_x}{k}$. Jos taas $s_x > k$, lemman 9 avulla saadaan $(-1)\binom{s_x}{k}\binom{s_x-k-1}{\alpha-k-1}$. Viimeinen tulos pätee itse asiassa kaikissa luetelluissa tapauksissa: jos $s_x < k$, $\binom{s_x}{k} = 0$, ja jos $s_x = k$, $\binom{s_x-k-1}{\alpha-k-1} = \binom{-1}{\alpha-k-1} = (-1)^{\alpha-k-1}$ ja siis $-\binom{s_x}{k}\binom{s_x-k-1}{\alpha-k-1} = (-1)^{\alpha+k}\binom{s_x}{k}$.

Tarkastellaan hyperverkkotermiä summana alkioiden $x \in U$ suhteen. Olkoon jonon $\mathcal{A}_x$ indeksijoukko I_x; siis $j \in I_x$, jos ja vain jos $x \in A_j$, ja $J \subseteq I_x$, jos ja vain jos $x \in \bigcap_{j\in J} A_j$. Tällä merkinnällä

$$S_{(H)} = \sum_{J\in E}\Big|\bigcap_{j\in J} A_j\Big| = \sum_{J\in E}\sum\Big(1 \;\Big|\; x \in \bigcap_{j\in J} A_j\Big) = \sum_{x\in U}\sum\Big(1 \;\Big|\; J \in E \cap I_x^{(\alpha)}\Big).$$

Niillä $x \in U$, joilla $s_x < \alpha$, edellisen summan termi on selvästikin 0. Jos taas $s_x \geq \alpha$, lauseen ehdosta saadaan termille arvio

$$\sum\left(1 \mid J \in E \cap I_x^{(\alpha)}\right) \leq \binom{s_x}{k}\binom{s_x-k-1}{\alpha-k-1}.$$

Asetetaan $f(x) = 0$, kun $s_x < \alpha$, ja $f(x) = \binom{s_x}{k}\binom{s_x-k-1}{\alpha-k-1}$, kun $s_x \geq \alpha$. Epäyhtälön (27) oikealle puolelle saadaan arvio

$$(-1)^{\alpha+k}\sum_{j=k}^{\alpha-1}(-1)^{j+k}\binom{j}{k}S_j + S_{(H)} \leq \sum_{x \in U}\left[f(x) - \binom{s_x}{k}\binom{s_x-k-1}{\alpha-k-1}\right].$$

Hakasulkujen sisään kirjoitettu lauseke on 0, kun $s_x \geq \alpha$, ja $-\binom{s_x}{k}\binom{s_x-k-1}{\alpha-k-1}$, kun $s_x < \alpha$. Kun $s_x < k$, $\binom{s_x}{k} = 0$, ja kun $k < s_x < \alpha$, $\binom{s_x-k-1}{\alpha-k-1} = 0$. Jäljelle jää tapaus $s_x = k$ eli $x \in Q_k$. Silloin lauseke on $-\binom{-1}{\alpha-k-1} = (-1)^{\alpha+k}$. Tästä seuraa väite. $\square$

Lukua P_k koskeva epäyhtälö on jälleen vaikeampi todistaa. Tässäkin tarvitaan hankalan näköistä lemmaa, joka todistetaan liitteessä A.

Lause 23. *Olkoot $k, \alpha \in \mathbb{Z}^+$ luvut, joille $k < \alpha$. Olkoon $H = \langle [n], E\rangle$ α-uniforminen hyperverkko, joka toteuttaa seuraavan ominaisuuden: minkä tahansa joukon $I \subseteq [n]$ indusoimassa alihyperverkossa on enintään*

$$\sum_{j=\alpha}^{|I|}\binom{j-1}{k-1}\binom{j-k-1}{\alpha-k-1}$$

kaarta. Silloin

$$(-1)^{\alpha+k}|P_k| \geq (-1)^{\alpha+k}\sum_{j=k}^{\alpha-1}(-1)^{j+k}\binom{j-1}{k-1}S_j + S_{(H)}. \tag{29}$$

Todistus. Kirjoitetaan katkaistu inkluusio-ekskluusio -kaava summana alkioiden $x \in U$ suhteen lemman 8 avulla,

$$(-1)^{\alpha+k}\sum_{j=k}^{\alpha-1}(-1)^{j+k}\binom{j-1}{k-1}S_j = \sum_{x \in U}\sum_{j=k}^{\alpha-1}(-1)^{j+\alpha}\binom{j-1}{k-1}\binom{s_x}{j}. \tag{30}$$

Samalla tavalla saadaan hyperverkkotermistä

$$S_{(H)} = \sum_{x \in U}\sum\left(1 \;\middle|\; J \in E \cap I_x^{(\alpha)}\right), \tag{31}$$

missä I_x on jonon $\mathcal{A}_x$ indeksijoukko.

Lasketaan yhteen summat (30) ja (31):

$$\begin{aligned}
&\sum_{x\in U}\left[\sum_{j=k}^{\alpha-1}(-1)^{j+\alpha}\binom{j-1}{k-1}\binom{s_x}{j}+\left|E\cap I_x^{(\alpha)}\right|\right]\\
\leq&\sum_{x\in U}\left[\sum_{j=k}^{\alpha-1}(-1)^{j+\alpha}\binom{j-1}{k-1}\binom{s_x}{j}+\sum_{j=\alpha}^{s_x}\binom{j-1}{k-1}\binom{j-k-1}{\alpha-k-1}\right]\\
=&\sum_{x\in U}\begin{cases}0, & \text{jos } s_x<\alpha,\\ (-1)^{k+\alpha}, & \text{jos } s_x\geq\alpha\end{cases}\\
=&(-1)^{\alpha+k}\left|P_k\right|.
\end{aligned}$$

Tässä käytettiin ensin hyperverkolta H vaadittua ehtoa ja sitten lemmaa 30 (sivulla 60).

$\square$

Edellisten lauseiden ehdot hyperverkoille näyttävät jossain määrin monimutkaisilta, mutta se johtuu rajojen optimaalisuudesta. Lauseiden 21, 22 ja 23 todistuksissa käytetään olennaisesti vain yhtä epäyhtälöä, joka tulee suoraan oletuksesta. Jos $H = \langle [n], E\rangle$ ei toteuta oletusta, vaan sillä on liian suuri alihyperverkko $\langle I, E\cap I^{(\alpha)}\rangle$, alkioita $x\in\bigcap_{j\in I}A_j$ vastaa liian suuri termi. Silloin voidaan konstruoida sellaiset U ja $\mathcal{A}$, että $U=\left(\bigcap_{j\in I}A_j\right)\setminus\left(\bigcup_{j\in[n]\setminus I}A_j\right)$, jolloin summan kaikki termit ovat liian suuria.

6 Sovellusongelman ratkaisuyrityksiä

Tässä luvussa palataan luvun 3 lopussa esitettyyn ongelmaan. Luvussa 6.1 kerrataan ensin ongelma hieman abstraktimmin kuin aiemmin, sitten luvussa 6.2 esitetään joitakin aikaisemmista epäyhtälöistä mieleen tulevia ratkaisuja. Kappale 6.3 käsittelee erästä ratkaisuyritystä, joka osoitetaan liian suoraviivaiseksi kappaleissa 6.4 ja 6.5. Kappaleessa 6.6 tarkastellaan vielä joitakin muita vaihtoehtoja ja esitetään mahdollisia jatkosuuntia.

Sekä ongelma että kappaleiden 6.4 ja 6.5 konjektuurit ovat Heikki Mannilan esittämiä, samoin lause 27.

6.1 Ongelman kuvaus

Kiinnitetään jälleen perusjoukko U ja sen osajoukkojen jono $\mathcal{A}$. Yhdisteen $\bigcup \mathcal{A}$ kokoa halutaan arvioida, kun tunnetaan leikkausten $\bigcap \mathcal{B}$ kokoja joillakin mutta ei kaikilla $\mathcal{B} \subseteq \mathcal{A}$. Tarkemmin sanottuna näitä kokoja on laskettu sellaisella algoritmilla, joka ottaa parametrina rajan $\rho \in [0, 1]$, valitsee ne osajonot $\mathcal{B} \subseteq \mathcal{A}$, joille $|\bigcap \mathcal{B}| \geq \rho\,|U|$, ja laskee näitä osajonoja vastaavien leikkausten koot $|\bigcap \mathcal{B}|$. Tällöin tunnettuja leikkauksia vastaavat osajonot muodostavat seuraavassa mielessä säännöllisen perheen: jos $|\bigcap \mathcal{B}|$ tunnetaan, tunnetaan myös $|\bigcap \mathcal{B}'|$ kaikilla $\mathcal{B}' \subseteq \mathcal{B}$.

Mainitun algoritmin sivutuotteina tunnetaan myös sellaisten leikkausten koot, jotka ovat aivan rajalla: jos $\mathcal{B} \subseteq \mathcal{A}$ on sellainen osajono, että $|\bigcap \mathcal{B}| < \rho\,|U|$ mutta jokaiselle $\mathcal{B}' \subsetneq \mathcal{B}$ pätee $|\bigcap \mathcal{B}'| \geq \rho\,|U|$, algoritmi laskee myös luvun $|\bigcap \mathcal{B}|$.

Jonon $\mathcal{A}$ joukkojen koot voisivat sattumalta olla sillä tavalla homogeenisia, että ehto $|\bigcap \mathcal{B}| \geq \rho\,|U|$ olisi voimassa täsmälleen niille jonoille $\mathcal{B} \subseteq \mathcal{A}$, joille $|\mathcal{B}| \leq k$ jollakin $k \in \mathbb{N}$. Tässä tapauksessa voidaan arvioida yhdisteen kokoa laskemalla inkluusio-ekskluusio -summa tunnettujen jonojen $\mathcal{B}$ suhteen,

$$|\bigcup \mathcal{A}| \approx b(\mathcal{A}, k) = \sum \Big((-1)^{|\mathcal{B}|}\,|\bigcap \mathcal{B}| \;\Big|\; |\mathcal{B}| \leq k \Big).$$

Arvion virheelle saadaan raja Bonferronin epäyhtälöstä:

$$|\#\bigcup \mathcal{A} - b(\mathcal{A}, k)| \leq \sum \Big(|\bigcap \mathcal{B}| \;\Big|\; |\mathcal{B}| = k + 1 \Big).$$

Tilanne on siinä mielessä hyvä, että algoritmin sivutuotteina lasketaan juuri jälkimmäiseen summaan tarvittavat leikkausten koot. Tavoitteena on yleistää edellistä päättelyä niin, ettei homogeenisuusoletusta tarvita, vaan algoritmin tuottamasta tiedosta saadaan joka tapauksessa sekä arvio luvulle $|\bigcup \mathcal{A}|$ että raja arvion virheelle.

Edellinen ongelman kuvaus on abstraktimpi kuin luvussa 3 esitetty. Seuraavassa kerrotaan, miten nämä ongelmat liittyvät toisiinsa; lukija, joka jätti luvun 3 väliin, voi siirtyä kappaleeseen 6.2.

Luvun 3 lopussa tarkasteltiin propositiosääntöä

$$\varphi \quad = \quad B \Rightarrow A_0 \vee A_1 \vee \cdots \vee A_{n-1},$$

missä $B, A_0, \ldots, A_{n-1}$ ovat relaatioskeeman S alkioita. Säännön kattavuuden tuntemiseksi halutaan laskea inkluusio-ekskluusio -summa

$$\sum\Big((-1)^{|J|+1}\operatorname{freq}(A_J)\;\Big|\;\emptyset \neq J \subseteq [n]\Big), \tag{32}$$

missä $A_J = \{\, A_j \mid j \in J \,\} \cup \{B\}$. Kuitenkin joukkojen A_J tiheyksistä tunnetaan vain ne, jotka ylittävät rajan $\rho \in [0,1]$, ja ne, jotka lasketaan sivutuotteina.

Tässä konkreettisessa tapauksessa universumia U vastaa skeemaa S noudattava relaatio R. Koska relaatio on määritelty jonoksi (eikä joukoksi), valitaan U:ksi R:n indeksijoukko. Kullakin $x \in U$ merkitään r_x:llä sitä R:n alkiota, jonka indeksi on x. Kutakin S:n hahmoa A_J, $J \in [n]$, vastaa nyt U:n osajoukko $U_J = \{\, x \in U \mid A_J \subseteq r_x \,\}$. Valitaan jonoon $\mathcal{A}$ yksiöitä $\{j\} \subseteq [n]$ vastaavat indeksijoukot, $\mathcal{A} = (\, U_{\{j\}} \mid j \in [n]\,)$. Koska $\operatorname{freq}(A_J) = |U_J| \,/\, |U|$, summalle (32) saadaan tuloksia arvioimalla summaa

$$|\bigcup \mathcal{A}| = \sum\Big((-1)^{|J|+1}\,|U_J|\;\Big|\;J \subseteq [n]\Big).$$

6.2 Aikaisempien epäyhtälöiden soveltaminen

Tunnettujen leikkausten joukossa ovat joka tapauksessa (kokokriteerin nojalla tai sivutuotteina) yhden joukon leikkaukset $\bigcap\{A_j\} = A_j$ kaikilla $j \in [n]$. Siis tasosumma S_1 osataan laskea, ja siitä saadaan yhdisteen koolle yläraja, joka on luultavasti melko huono. Jos tunnettujen leikkausten joukko sisältää myös kaikki kahden joukon leikkaukset $\bigcap\{A_i, A_j\}$, $\{i,j\} \in [n]^{(2)}$, Bonferronin epäyhtälöstä saadaan myös alaraja $S_1 - S_2$. Näin voidaan jatkaa niin pitkälle kuin kokonaisia tasoja riittää.

Edellinen ajatus kaatuu siihen, että jo toisessa tasosummassa S_2 on noin $n^2/2$ termiä, ja seuraavissa vielä enemmän (suurilla n). Ei vaikuta kovin todennäköiseltä, että kaikki nämä termit olisivat laskettujen leikkausten joukossa; jos ne ovat, raja ρ on luultavasti valittu liian pieneksi.

Jos tunnettaisiin kokonaisia tasosummia, kannattaisi luultavasti käyttää esimerkiksi kirjassa [GS96] johdettuja lineaariseen ohjelmointiin perustuvia epäyhtälöitä.

Luvussa 5 esitellyt verkkoepäyhtälöt tarjoavat toisen mahdollisuuden. Algoritmi voisi sisältää kuvauksia erilaisista verkoista ja yrittää sovittaa niitä tunnettuihin leikkauksiin. Tämän menetelmän hankaluutena näyttää olevan se, ettei ennalta osata ennustaa tunnettujen leikkausten joukkoa. Jos algoritmi tunnistaa vain muutamia verkkoja, onnistumisen todennäköisyys on pieni suunnilleen samasta syystä kuin toivottaessa, että riittävän moni symmetrinen tasosumma saadaan laskettua. Jos taas pyritään tunnistamaan monta eri vaihtoehtoa, algoritmi on joko hidas tai sen kirjoittaminen on vaikea tehtävä.

Yksi mahdollisuus soveltaa verkkoepäyhtälöitä on sellainen, että lisätään tunnettujen leikkausten joukkoon jokin pieni määrä ylimääräisiä leikkauksia, jotta saadaan jokin hyvä verkko. Jos jo siinä vaiheessa, kun tietokantaa käydään läpi leikkausten laskemiseksi, voitaisiin tunnistaa otolliset tilanteet ja ottaa mukaan muutama ylimääräinen leikkaus, ei tarvittaisi ylimääräisiä läpikäyntejä eikä aikaa välttämättä kuluisi merkittävästi enempää.

Toisaalta voitaisiin edetä toiseen suuntaan: tarkastellaan tunnettuihin leikkauksiin liittyvää verkkoa ja yritetään todistaa sille sopivia ominaisuuksia. Kirjassa [GS96], sivuilla 38–45, esitellään tämänsuuntaista ajatusta.

Hyperverkkoihin perustuvissa epäyhtälöissä näyttää olevan se ongelma, että niiden käyttämiseksi pitäisi tuntea jokin määrä kokonaisia tasoja. Jos niitä tunnetaan, seuraavalta tasolta riittää valita jokin riittävän harva hyperverkko. Optimaalisen hyperverkon valintakaan ei ole aivan mitätön ongelma, vaan artikkeli [Gra93] omistaa sen pohtimiselle yhden luvun ja antaa vain algoritmin, joka etsii eräässä harvojen α-uniformisten hyperverkkojen osajoukossa optimaalisen hyperverkon.

6.3 Osittainen inkluusio-ekskluusio

Tunnettujen leikkausten joukko on siinä mielessä säännöllinen, että jos $|\bigcap \mathcal{B}|$ tunnetaan, tunnetaan myös $|\bigcap \mathcal{B}'|$ kaikilla $\mathcal{B}' \subseteq \mathcal{B}$. Seuraavassa esitetään tästä ajatuksesta kumpuavia käsitteitä ja kappaleissa 6.4 ja 6.5 niihin perustuvia ratkaisuyrityksiä.

Määritelmä 24. Olkoon I äärellinen joukko. Perhettä $\mathcal{J} \subseteq \mathcal{P}(I)$ sanotaan *alaspäin suljetuksi*, jos kaikilla $J' \subseteq J \in \mathcal{J}$ pätee $J' \in \mathcal{J}$.

Jos $\mathcal{J} \subseteq \mathcal{P}(I)$ on alaspäin suljettu perhe, sen *yläraja* $\partial\mathcal{J} \subseteq \mathcal{P}(I)$ sisältää täsmälleen ne joukot J, jotka eivät kuulu perheeseen $\mathcal{J}$ mutta joiden jokainen aito osajoukko kuuluu siihen. Toisin sanoen

$$\partial\mathcal{J} = \{\, J \subseteq I \mid J' \in \mathcal{J} \text{ kaikilla } J' \subsetneq J \,\} \setminus \mathcal{J}.$$

Alaspäin suljetun perheen voi helposti määritellä maksimaalisten joukkojen avulla. Jos nimittäin $J \in \mathcal{J}$, missä $\mathcal{J}$ on alaspäin suljettu, niin välttämättä $\mathcal{P}(J) \subseteq \mathcal{J}$. Siis $\mathcal{J}$ on esitettävissä yhdisteenä $\bigcup_{\ell \in L} \mathcal{P}(J_\ell)$, missä $\{\, J_\ell \mid \ell \in L \,\}$ on maksimaalisten joukkojen perhe. Toisaalta jokainen tällainen yhdiste on selvästi alaspäin suljettu.

Esimerkki 9. Erikoistapauksena maksimaalisia joukkoja voisivat olla kaikki I:n k-joukot. Määritellään

$$\mathcal{J}_k = \bigcup \{\, \mathcal{P}(J) \mid J \in I^{(k)} \,\}.$$

Jos joukossa $J \subseteq I$ on enintään k alkiota, niin varmasti $J \in \mathcal{J}_k$. Jos taas $|J| > k$, $J \notin \mathcal{J}_k$. Saadaan siis toinen esitys tasojoukkojen yhdisteenä

$$\mathcal{J}_k = \{\, J \subseteq I \mid \#J \leq k \,\} = \bigcup \{\, I^{(\ell)} \mid 0 \leq \ell \leq k \,\}.$$

Perheessä $\partial\mathcal{J}_k$ ovat siis täsmälleen ne joukot, joiden jokainen aito osajoukko on enintään k-alkioinen mutta jotka eivät kuulu $\mathcal{J}$:hin. Näitä ovat tietysti I:n $k+1$-alkioiset joukot,

$$\partial\mathcal{J}_k = I^{(k+1)}.$$

Alaspäisen suljettuuden ja ylärajan käsitteet liittyvät ongelmaan sitä kautta, että I:tä käytetään jonkin joukkojonon (A_i) indeksointiin ja sen osajoukkoihin J liitetään leikkauksia $\bigcap(A_i \mid i \in J)$. Tunnettuihin leikkauksiin liittyvät indeksijoukot muodostavat alaspäin suljetun perheen, ja sivutuotteena tunnettujen leikkausten indeksijoukot muodostavat edellisen perheen ylärajan. Leikkausten ja indeksijoukkojen yhteydestä puhumiseksi asetetaan seuraava määritelmä.

Määritelmä 25. Olkoon U äärellinen joukko ja $\mathcal{A} = (A_i \mid i \in [n])$ jono sen osajoukkoja. Kun $J \subseteq [n]$, merkitään $X(\mathcal{A}, J) = |\bigcap(A_i \mid i \in J)|$.

Nyt voidaan sanoa, että tunnetaan jono $(X(\mathcal{A}, J) \mid J \in \mathcal{J})$, missä $\mathcal{J} \in \mathcal{P}([n])$ on alaspäin suljettu, ja että sivutuotteena tunnetaan myös jono $(X(\mathcal{A}, J) \mid J \in \partial\mathcal{J})$.

Yksi luonnollinen arvaus yhdisteen $\bigcup \mathcal{A}$ kooksi on inkluusio-ekskluusio -summa (6) (sivulla 19) laskettuna vain tunnettujen joukkojen suhteen:

$$b(\mathcal{A}, \mathcal{J}) = \sum \Big((-1)^{|J|+1} X(\mathcal{A}, J) \;\Big|\; \emptyset \neq J \in \mathcal{J} \Big). \tag{33}$$

Tämä on yleistys Bonferronin epäyhtälön (lause 11 sivulla 28) summasta $b(\mathcal{A}, k)$. Esimerkin 9 tapauksessa kyseessä on juuri sama summa: $b(\mathcal{A}, \mathcal{J}_k) = b(\mathcal{A}, k)$. Määritellään myös vastaava virhe

$$d(\mathcal{A}, \mathcal{J}) = |\# \bigcup \mathcal{A} - b(\mathcal{A}, \mathcal{J})| \,. \tag{34}$$

Kiinnostava kysymys on, mitä tästä virheestä osataan sanoa. Esimerkin 9 tapauksessa Bonferronin epäyhtälö antaa vastauksen $d(\mathcal{A}, \mathcal{J}_k) = d(\mathcal{A}, k) \leq S_{k+1}$. Seuraavissa kahdessa kappaleessa tutkitaan kahta yritystä yleistää Bonferronin epäyhtälöä tähän sopivaksi.

6.4 Ensimmäinen yritys: yläraja

Seuraava konjektuuri pyrkii yleistämään Bonferronin epäyhtälöä käsillä olevaan tapaukseen. Korvataan virheen rajana käytetty summa S_{k+1} ylärajan koolla:

$$d(\mathcal{A}, \mathcal{J}) \overset{?}{\leq} \sum \Big(X(\mathcal{A}, J) \;\Big|\; J \in \partial\mathcal{J} \Big). \tag{35}$$

Seuraava lause osoittaa, että epäyhtälö (35) ei pidä paikkaansa. Siinä konstruoidaan melko monimutkainen vastaesimerkki, jota käytetään myös seuraavan konjektuurin vastaesimerkin pohjana.

Lause 24. *Epäyhtälö* (35) *ei päde kaikilla* U, $\mathcal{A}$ *ja* $\mathcal{J}$.

Todistus. Asetetaan $U = \{x\}$, $n = 10$ ja $A_i = U$ kaikilla $i \in [10]$. Selvästi $|\bigcup \mathcal{A}| = |U| = 1$.

Määritellään alaspäin suljettu jono $\mathcal{J}$ komplementtinsa minimaalisten joukkojen avulla. Olkoon $\mathcal{T} \subseteq \mathcal{P}([10])$ seuraava perhe:

$$\mathcal{T} = \{ \{4,5,6,7,8,9\}, \{1,2,3,7,8,9\}, \{0,2,3,5,6,9\}, \{0,1,3,4,6,8\}, \{0,1,2,4,5,7\} \}.$$

0	1	2	3	4	5	6	7	8	9
				•	•	•	•	•	•
	•	•	•				•	•	•
•		•	•		•	•			•
•	•		•	•		•		•	
•	•	•		•	•		•		

Taulukko 4: Perhe $\mathcal{T}$

Perhettä $\mathcal{T}$ havainnollistaa taulukko 4, jonka viisi vaakariviä vastaavat $\mathcal{T}$:n joukkoja ja kymmenen pystyriviä indeksijoukon [10] alkioita. Tästä esityksestä nähdään välittömästi, että jokaista indeksiä vastaa tasan kaksi $\mathcal{T}$:n joukkoa, joihin se ei kuulu, ja että millään kahdella indeksillä nämä joukot eivät ole samat. Toisaalta $\mathcal{T}$:llä on kymmenen kaksialkioista osaperhettä, joten jokaista sellaista vastaa tasan yksi indeksi, joka ei kuulu kumpaankaan osaperheen joukkoon.

Määritellään seuraavaksi alaspäin suljettu perhe $\mathcal{J}$ perheen $\mathcal{T}$ avulla. Valitaan siihen tasan ne joukot $J \subseteq [10]$, joilla ei ole mitään $\mathcal{T}$:n joukkoa osajoukkonaan, toisin sanottuna

$$\mathcal{J} = \{\, J \subseteq [10] \mid J \not\supseteq T \text{ kaikilla } T \in \mathcal{T} \,\}.$$

Koska käytetty valintaehto periytyy osajoukoille, $\mathcal{J}$ on alaspäin suljettu.

Jos S on jonkin $\mathcal{T}$:n joukon aito osajoukko, $S \subsetneq T^* \in \mathcal{T}$, sillä on varmasti ominaisuus $S \not\supseteq T$ kaikilla $T \in \mathcal{T}$ jo senkin takia, että $|S| \leq 5$. Silloin $S \in \mathcal{J}$, joten $\mathcal{T} \subseteq \partial\mathcal{J}$. Olkoon nyt $R \in \partial\mathcal{J}$. Koska $R \notin \mathcal{J}$, $\mathcal{J}$:n määritelmän nojalla pätee $R \supseteq T$ jollain $T \in \mathcal{T}$. Jos T on R:n aito osajoukko, saadaan ristiriita ylärajan määritelmän kanssa, koska $T \notin \mathcal{J}$. Siis $R = T \in \mathcal{T}$, joten $\partial\mathcal{J} \subseteq \mathcal{T}$. Saatiin tulos $\partial\mathcal{J} = \mathcal{T}$.

Lasketaan nyt $b(\mathcal{A}, \mathcal{J})$. Kirjoitetaan kaava (33) tasoittain,

$$b(\mathcal{A}, \mathcal{J}) = \sum_{k=1}^{10} (-1)^{k+1} S_{\mathcal{J},k},$$

missä

$$S_{\mathcal{J},k} = \sum \Big(X(\mathcal{A}, J) \;\Big|\; J \in \mathcal{J}, |J| = k \Big). \tag{36}$$

Koska kaikki alle kuusialkioiset indeksijoukot ovat $\mathcal{J}$:ssä, $S_{\mathcal{J},k} = \big|[10]^{(k)}\big| = \binom{10}{k}$, kun $k \in \{1, 2, 3, 4, 5\}$. Kuusialkioisista joukoista kaikki paitsi $\mathcal{T}$:n joukot ovat $\mathcal{J}$:ssä, joten $S_{\mathcal{J},6} = \big|[10]^{(6)} \setminus \mathcal{T}\big| = 205$.

Lasketaan $S_{\mathcal{J},7}$ inkluusio-ekskluusio -periaatteella. Kukin $T \in \mathcal{T}$ on neljän seitsenalkioisen $\mathcal{P}([10])$:n joukon osajoukko, nimittäin kutakin $i \in [10] \setminus T$ vastaa joukko $T \cup \{i\}$. Toisaalta mikään seitsenalkioinen indeksijoukko ei sisällä osajoukkonaan enempää kuin yhden $\mathcal{T}$:n joukon: kuten edellä todettiin, kaikki kahden eri $\mathcal{T}$:n joukon yhdisteet ovat yhdeksänalkioisia. Siis inkluusio-ekskluusio -summa päättyy ensimmäiseen tasoon, ja $S_{\mathcal{J},7} = \binom{10}{7} - 5 \cdot 4 = 100$.

Vastaavasti voidaan päätellä, että $S_{\mathcal{J},8} = \binom{10}{8} - 5 \cdot 6 = 15$. Kullakin $\mathcal{T}$:n joukolla on nimittäin $\binom{4}{2} = 6$ kahdeksanalkioista ylijoukkoa $\mathcal{P}([10])$:ssä.

Koska jokaisella yhdeksänalkioisella indeksijoukolla on jokin $\mathcal{T}$:n joukko osajoukkonaan, $S_{\mathcal{J},9} = S_{\mathcal{J},10} = 0$. Saatiin tulos $b(\mathcal{A}, \mathcal{J}) = \sum_k (-1)^{k+1} S_{\mathcal{J},k} = 10 - 45 + 120 - 210 + 252 - 205 + 100 - 15 + 0 - 0 = 7$. Yhdisteen todellinen koko on $|\bigcup \mathcal{A}| = |U| = 1$, joten arvion virhe on $d(\mathcal{A}, \mathcal{J}) = 6$. Ylärajassa on viisi joukkoa J, joista jokaiselle pätee $X(\mathcal{A}, J) = 1$. Epäyhtälön (35) oikean puolen summa on siis 5, joten epäyhtälö on osoitettu vääräksi. □

Huomautus 10. Vastaesimerkin olemassaolo voisi toki johtua siitä, että konjektuuri on vain hieman löysästi muotoiltu. Ensinnäkin vastaesimerkki koskee vain yksiöitä $U = \{x\}$, jotka eivät ole todellisuudessa kovin kiinnostavia. Epäyhtälö (35) on kuitenkin sellaista muotoa, että Rényin lause 13 pätee, joten vastaesimerkkiä riittää etsiä yksiöavaruuksista.

Varsinaisessa ongelmassa perheen $\mathcal{J}$ määrää kuitenkin ρ-kattavuus. Voisihan olla, ettei tällainen $\mathcal{J}$ esiinny koskaan todellisuudessa. Tutkitaan nyt, miten mielivaltainen $\mathcal{J}$ saadaan tuotettua ρ-kattavuuden määräämänä.

Pidetään aluksi joukot A_i samoina. Lisätään joukkoon U sekä sellaisia alkioita, jotka kuuluvat kaikkiin A_i, että sellaisia, jotka eivät kuulu mihinkään; näin saadaan jokaisen joukon A_i osuus $|A_i| / |U|$ mielivaltaisen lähelle suhdetta ρ. Ryhdytään sitten muuttamaan joukkoja A_i lisäämällä U:hun vielä ylimääräisiä alkioita. Lisätään kutakin $\mathcal{J}$:n maksimaalista joukkoa J vastaaviin joukkoihin A_i, $i \in J$, alkio x_J. Näin saadaan aikaan tilanne, jossa $X(\mathcal{A}, J) \geq \rho |U|$, jos ja vain jos $J \in \mathcal{J}$. Koska sekä rajan $\rho |U|$ ylittävien että sen alittavien leikkausten koot saadaan mielivaltaisen lähelle rajaa ja $\mathcal{J}$ on vain äärellinen, vastaesimerkin arviot saadaan pätemään niin tarkasti, että lopputulos on edelleen voimassa.

6.5 Seuraava yritys: vakio kertaa yläraja

Epäyhtälö (35) osoittautui liian yksinkertaiseksi. Esimerkki on kuitenkin jossain määrin monimutkainen, ja silläkin saatiin aikaan vain suhteellisen pieni virhe. Saattaisi olla, että käytetty ylärajan käsite on mielekäs, kun epäyhtälöä väljennetään hiukan. Tarkastellaan seuraavaksi epäyhtälöä

$$d(\mathcal{A}, \mathcal{J}) \overset{?}{\leq} c \sum \Big(X(\mathcal{A}, J) \;\Big|\; J \in \partial \mathcal{J} \Big), \tag{37}$$

missä c on jokin positiivinen vakio, joka ei riipu U:sta, $\mathcal{A}$:sta eikä $\mathcal{J}$:stä. Seuraavassa yleistetään edellistä vastaesimerkkiä niin, että nähdään tämäkin epäyhtälö pätemättömäksi.

Olkoon edelleen $U = \{x\}$. Olkoon $m \in \mathbb{Z}^+$, $m \geq 2$, ja $n = \binom{m}{2} = m(m-1)/2$. Asetetaan jälleen $A_i = U$ kaikilla $i \in [n]$.

Joukolla $[m]$ on n kaksialkioista osajoukkoa; numeroidaan nämä osajoukot jollain kanonisella tavalla niin, että kutakin 2-joukkoa $\{r,s\} \in [m]^{(2)}$ vastaa yksikäsitteinen indeksi $i_{\{r,s\}} \in [n]$. Konstruoidaan perhe $\mathcal{T} \subseteq \mathcal{P}([n])$, $\mathcal{T} = \{\, T_j \mid j \in [m]\,\}$, säännöllä

$$T_j = \{\, i \in [n] \mid \nexists k \in [m] \setminus \{j\} : i = i_{\{j,k\}}\,\}.$$

Toisin sanoen joukkoon T_j kuuluu jokainen luku $i \in [n]$, jota edellisessä numeroinnissa vastaava 2-joukko ei sisällä lukua j. Heti havaitaan, että $|T_j| = n - m + 1$. Osoitetaan nyt, että perheellä $\mathcal{T}$ on vastaava ominaisuus kuin edellisen kappaleen erikoistapauksessa.

Lemma 25. *Kaikilla $i \in [n]$ on olemassa sellaiset $r, s \in [m]$, että $T_r \cup T_s = [n] \setminus \{i\}$. Käänteisesti kaikille $r, s \in [m]$, $r \neq s$, pätee $T_r \cup T_s = [n] \setminus \{i_{\{r,s\}}\}$.*

Todistus. Numeroinnin bijektiivisyyden perusteella kaikkia indeksejä $i \in [n]$ vastaavat sellaiset $r, s \in [m]$, että $i = i_{\{r,s\}}$. Joukkojen T_j määritelmän nojalla $i_{\{r,s\}} \notin T_r \cup T_s$. Jos toisaalta $i' \neq i$, niin $i' = i_{\{r',s'\}}$, missä $\{r', s'\} \neq \{r, s\}$, joten $i' \in T_r \cup T_s$. Siis $T_r \cup T_s = I \setminus \{i\}$.

Väitteen toisen osan todistamiseksi todetaan ensin, että $i_{\{r,s\}} \notin T_r \cup T_s$. Jos nyt $i' \notin T_r \cup T_s$, kirjoitetaan $i' = i_{\{r',s'\}}$. Koska $i' \notin T_r$, $r \in \{r', s'\}$, ja koska $i' \notin T_s$, myös $s \in \{r', s'\}$. Mutta silloin $\{r, s\} = \{r', s'\}$ ja siis $i' = i_{\{r,s\}}$. □

Määritellään nyt $\mathcal{J}$ ehdolla

$$\mathcal{J} = \{\, J \subseteq [n] \mid J \not\supseteq T \text{ kaikilla } T \in \mathcal{T}\,\}.$$

Samaan tapaan kuin edellisessä kappaleessa voidaan päätellä, että $\mathcal{J}$ on alaspäin suljettu ja $\partial \mathcal{J} = \mathcal{T}$.

Lasketaan osittainen inkluusio-ekskluusio -summa (33) tasoittain,

$$b(\mathcal{A}, \mathcal{J}) = \sum_{k=1}^{n} (-1)^{k+1} S_{\mathcal{J},k},$$

missä tasosumma $S_{\mathcal{J},k}$ on yhtälön (36) määrittelemä. Kaikilla $k \leq n - m$ pätee $S_{\mathcal{J},k} = \binom{n}{k}$, koska mikään joukko, jossa on korkeintaan $n - m$ alkiota, ei sisällä osajoukkonaan mitään T_j:tä. Lemman 25 nojalla $S_{\mathcal{J},n-1} = 0$, ja samoin selvästi $S_{\mathcal{J},n} = 0$.

Olkoon nyt $n - m < k < n - 1$. Koska kaikissa kahden $\mathcal{T}$:n joukon yhdisteissä on $n - 1$ alkiota, jokainen $J \in [n]^{(k)}$ sisältää osajoukkonaan enintään yhden $\mathcal{T}$:n joukon. Kullakin j sellaisia $J \in [n]^{(k)}$, joille $T_j \subseteq J$, on $\binom{m-1}{k-n+m-1} = \binom{m-1}{n-k}$. Nimittäin joukossa J on välttämättä jokainen joukon T_j alkioista, joita on $n - m + 1$, ja $k - n + m - 1$ muuta alkiota, jotka voidaan valita niistä $n - (n - m + 1) = m - 1$ alkiosta, jotka eivät kuulu joukkoon T_j. Kaikkiaan $S_{\mathcal{J},k} = \binom{n}{k} - m\binom{m-1}{n-k}$.

Nyt voidaan laskea osittainen inkluusio-ekskluusio -summa.

$$
\begin{aligned}
b(\mathcal{A}, \mathcal{J}) &= \sum_{k=1}^{n} (-1)^{k+1} S_{\mathcal{J},k} \\
&= \sum_{k=1}^{n-m} (-1)^{k+1} \binom{n}{k} + \sum_{k=n-m+1}^{n-2} (-1)^{k+1} \left[\binom{n}{k} - m \binom{m-1}{n-k} \right] \\
&= \sum_{k=1}^{n-2} (-1)^{k+1} \binom{n}{k} - m \sum_{k=n-m+1}^{n-2} (-1)^{k+1} \binom{m-1}{n-k} \\
&= -(1-1)^n - (-1)^{0+1} \binom{n}{0} - (-1)^n \binom{n}{n-1} - (-1)^{n+1} \binom{n}{n} \\
&\qquad - m \sum_{\ell=2}^{m-1} (-1)^{n-\ell+1} \binom{m-1}{\ell} \\
&= 0 - (-1) - (-1)^n n - (-1)^{n+1} \\
&\qquad - (-1)^{n+1} m \left[(1-1)^{m-1} - (-1)^0 \binom{m-1}{0} - (-1)^1 \binom{m-1}{1} \right] \\
&= 1 + (-1)^n \Big(-n + 1 + m(0 - 1 + m - 1) \Big) \\
&= 1 + (-1)^n \big(-m(m-1)/2 + 1 + m(m-2) \big) = 1 + (-1)^n \big(m^2 - 3m + 2 \big)/2.
\end{aligned}
$$

Koska $|\bigcup \mathcal{A}| = |U| = 1$, osittaisen inkluusio-ekskluusion virhe on

$$
d(\mathcal{A}, \mathcal{J}) = \big(m^2 - 3m + 2 \big)/2.
$$

Ylärajassa on m joukkoa, joista jokaisen koko on 1, siis epäyhtälön (37) oikea puoli on cm. Selvästikin $d(\mathcal{A}, \mathcal{J})/m \to \infty$, kun $m \to \infty$. On todistettu seuraava lause.

Lause 26. *Epäyhtälö* (37) *ei päde millään vakiolla c kaikilla U, $\mathcal{A}$ ja $\mathcal{J}$.*

Huomautus 10 (sivulla 52) pätee yhtä lailla lauseeseen 26 kuin lauseeseen 24.

6.6 Muita mahdollisuuksia

Lause 26 osoittaa, että epäyhtälössä 37 on luovuttava siitä vaatimuksesta, että c on vakio. Jos perheen $\mathcal{J}$ perusteella pystyttäisiin laskemaan sellainen kohtuullisen pieni c, että epäyhtälö olisi tosi, ongelma olisi tyydyttävästi ratkaistu. Ratkaisulta olisi toivottavaa, että esimerkin 9 (sivulla 49) perheelle J_k saataisiin $c = 1$.

Seuraava lause ei anna hyvää ratkaisua, mutta siitä saadaan kuitenkin yläraja c:lle.

Lause 27. *Epäyhtälö (37) pätee, kun $c = 2^n$.*

Todistus. Kirjoitetaan virhe inkluusio-ekskluusio -periaatteen avulla summaksi

$$\begin{aligned}
d(\mathcal{A},\mathcal{J}) &= |\#\bigcup\mathcal{A} - b(\mathcal{A},\mathcal{J})| \\
&= \Big|\sum\big((-1)^{|J|+1}X(\mathcal{A},J) \mid \emptyset \neq J \subseteq [n]\big) - \sum\big((-1)^{|J|+1}X(\mathcal{A},J) \mid \emptyset \neq J \in \mathcal{J}\big)\Big| \\
&= \left|\sum\Big((-1)^{|J|+1}\sum\big(1 \mid x \in \bigcap_{j\in J} A_j\big) \;\Big|\; \emptyset \neq J \in \mathcal{P}([n]) \setminus \mathcal{J}\Big)\right| \\
&= \left|\sum_{x\in U}\sum\big((-1)^{|J|+1} \mid J \in \mathcal{P}([n]) \setminus \mathcal{J},\, x \in \bigcap_{j\in J} A_j\big)\right| \\
&\leq \sum_{x\in U}\sum\big(1 \mid J \in \mathcal{P}([n]) \setminus \mathcal{J},\, x \in \bigcap_{j\in J} A_j\big).
\end{aligned}$$

Kiinnitetään alkio $x \in U$ ja tarkastellaan siihen liittyvää termiä tässä summassa. Jos kaikille indeksijoukoille $J \in \mathcal{P}([n])$, joille $x \in \bigcap_{j\in J} A_j$, pätee $J \in \mathcal{J}$, alkioon x liittyvä termi on 0. Muussa tapauksessa olkoon

$$\mathcal{J}_x = \{\, J \in \mathcal{P}([n]) \setminus \mathcal{J} \mid x \in \bigcap_{j\in J} A_j \,\}$$

niiden indeksijoukkojen J perhe, jotka eivät kuulu perheeseen $\mathcal{J}$ mutta joille $x \in \bigcap_{j\in J} A_j$. Tässä perheessä on enintään $|\mathcal{P}([n])| = 2^n$ joukkoa, ja jokaista kohti summassa on termi 1. Siis alkiota $x \in U$ vastaava termi on joko 0 tai enintään 2^n.

Toisaalta joukossa $\mathcal{J}_x$ on osajoukkorelaation suhteen minimaalinen (ei välttämättä yksikäsitteinen) joukko $J_x \in \mathcal{J}_x$. Koska ominaisuus $x \in \bigcap_{j\in J} A_j$ periytyy joukon J osajoukoille, minimaalisuudesta seuraa, että $J_x \in \partial\mathcal{J}$. Koska jokaista sellaista $x \in U$, jota vastaa positiivinen termi, vastaa myös sellainen J_x, että $x \in X(\mathcal{A}, J_x)$, näitä x on enintään niin monta kuin joukoissa $X(\mathcal{A}, J)$, $J \in \partial\mathcal{J}$, alkioita. Siis

$$d(\mathcal{A},\mathcal{J}) \leq 2^n\Big|\bigcup\{\, X(\mathcal{A},J) \mid J \in \partial\mathcal{J} \,\}\Big| \leq 2^n \sum\Big(X(\mathcal{A},J) \;\Big|\; J \in \partial\mathcal{J}\Big),$$

mikä todistaa väitteen. □

Edellistä todistusta muuntamalla voitaisiin ehkä löytää jokin pienempi c, jolla (37) pätee. Todistus tuntuu jättävän aika paljon tilaa, koska koko joukkoa miinus ykkösiä arvioidaan ykkösillä. Toisaalta myös Bonferronin epäyhtälö on melko väljä, kuten esimerkiksi luvusta 5 nähdään.

Myös ylärajan käsitteen sopiva säätäminen voisi auttaa. Siihen voitaisiin esimerkiksi lisätä perheen $\mathcal{J}$ maksimaaliset joukot, tavallaan siis sen komplementin alaraja. Vastaesimerkki on hankala tapaus, koska sen ylärajassa on liian vähän joukkoja; ainakin näyttää vaikealta antaa sellainen esimerkki, jossa maksimaalisilla joukoilla vahvistettu yläraja olisi pieni. Toisaalta tätä ylärajaa vastaavan epäyhtälön erikoistapauksena ei saada tavallinen Bonferronin epäyhtälöä vaan vielä väljempi tulos.

Vielä yksi mahdollisuus liittyy lauseen 11 (sivulla 28) toiseen todistukseen. Sehän perustuu osajonohilan järjestämiseen pareiksi kuvan 2 (sivulla 24) tavalla. Jos tähän kuvaan piirretään vaakasuora viiva, se leikkaa enintään niin monta nuolta kuin seuraavalla tasolla on osajonojen leikkauksia. Katkaistun inkluusio-ekskluusion tapauksessa viiva ei ole vaakasuora, mutta samantapainen ajatus saattaisi toimia.

A Eräs binomikerrointen summa

Tässä liitteessä todistetaan kaksi teknistä lemmaa 29 ja 30. Ne liittyvät summiin, joiden termit ovat muotoa $(-1)^j\binom{j-1}{k-1}\binom{s}{j}$.

Lemmaa 29 tarvitaan lauseen 12 todistuksessa. Itse todistus on varsin suoraviivainen, mutta se vaatii paljon tarkistuslaskemista eikä sisällä mitään suuria oivalluksia. Seuraamisen helpottamiseksi esitetään ensin yksinkertaisempi lemma 28; tulos sinänsä ei ole kiinnostava, mutta todistusmenetelmä on samanlainen kuin lemmassa 29.

Lemma 28. *Määritellään funktio f kaavalla $f(n,j) = \binom{n}{j}$, missä $n \in \mathbb{N}$ ja $j \in \mathbb{Z}$, ja summafunktio F kaavalla*

$$F(n) = \sum_{j\in\mathbb{Z}} f(n,j).$$

Jälkimmäinen funktio toteuttaa kaikilla $n \in \mathbb{N}$ palautuskaavan

$$F(n+1) - 2F(n) = 0.$$

Todistus. Lemma olisi tietysti helppo todistaa havaitsemalla, että $F(n) = 2^n$, mutta käytetään esimerkin vuoksi hankalampaa menetelmää. Määritellään apufunktio g kaavoilla

$$g(n,j) = \begin{cases} \dfrac{j}{j-n-1}\dbinom{n}{j}, & j \neq n+1, \\ -1, & j = n+1. \end{cases}$$

Tätä määrittelyä motivoi palautuskaava

$$f(n+1,j) - 2f(n,j) = g(n,j+1) - g(n,j), \tag{38}$$

joka todistetaan seuraavassa kaikille $j \in \mathbb{Z}$, $n \in \mathbb{N}$. Kun lasketaan yhteen kaavan (38) kumpikin puoli kaikilla kokonaisluvuilla j, saadaan väite.

Kaavan (38) voi tarkistaa mekaanisesti. Jos $0 \le j \le n$, vasen puoli on

$$\begin{aligned}\binom{n+1}{j} - 2\binom{n}{j} &= \frac{(n+1)!}{j!(n-j+1)!} - \frac{2n!}{j!(n-j)!} \\ &= \frac{(n+1)! - 2n!(n-j+1)}{j!(n-j+1)!} = \frac{n!(2j-n-1)}{j!(n-j+1)!}.\end{aligned}$$

Jos $j < 0$ tai $j > n+1$, vasen puoli on 0. Jos $j = n+1$, saadaan 1.

Jos $0 \le j \le n-1$, oikea puoli on

$$\begin{aligned}\frac{j+1}{j-n}\binom{n}{j+1} - \frac{j}{j-n-1}\binom{n}{j} &= \frac{(j+1)n!}{(j-n)(j+1)!(n-j-1)!} - \frac{jn!}{(j-n-1)j!(n-j)!} \\ &= \frac{-n!}{j!(n-j)!} - \frac{-jn!}{j!(n-j+1)!} = \frac{n!(-(n-j+1)-(-j))}{j!(n-j+1)!} = \frac{n!(2j-n-1)}{j!(n-j+1)!}.\end{aligned}$$

Tapauksissa $j < 0$ ja $j > n+1$ oikea puoli on 0. Jos $j = n$, oikea puoli on $n-1$, ja jos $j = n+1$, oikea puoli on 1. Kaikissa tapauksissa vasen ja oikea puoli yhtyvät. $\square$

Todistuksen ydin on funktio g. Jos se ja yhtälö (38) tunnetaan, kaikki muu on niin helppoa, että sen voisi automatisoida tietokoneohjelmalla. Menetelmän tekee mielenkiintoiseksi se, että kirjassa [PWZ97] esitetään algoritmi, joka etsii eräille tämäntapaisille summille palautuskaavan sekä todistuksessa tarvittavan funktion g. Tämä funktio toimii eräänlaisena ratkaisun varmenteena: jos herää epäilys, että algoritmi on toiminut väärin, tulos on helppo tarkistaa varmenteen avulla.

Algoritmin toteuttava Maple-ohjelma löytyy verkosta[3]. Tämän ohjelman avulla saadaan seuraava tulos, jonka todistus on samanmuotoinen kuin edellinen.

Lemma 29. *Olkoot $k \leq \ell$ luonnollisia lukuja. Määritellään funktio f kaavalla $f(s,j) = (-1)^{j-k}\binom{j-1}{k-1}\binom{s}{j}$, kun $j \in \mathbb{Z}$, $s \in \mathbb{N}$, ja summafunktio F kaavalla*

$$F(s) = \sum_{j \geq \ell+1} f(s,j).$$

Jälkimmäinen funktio toteuttaa kaikilla $s \in \mathbb{N}$ palautuskaavan

$$(s-k+1)(F(s+1)-F(s)) = (-1)^{\ell+k+1}\binom{s}{\ell}\binom{\ell}{k-1}(\ell-k+1).$$

Todistus. Apufunktio g on tässä tapauksessa

$$g(s,j) = \begin{cases} \dfrac{(j-k)j}{j-s-1} f(s,j), & j \neq s+1, \\ (-1)^{s-k}(s-k+1)\dbinom{s}{k-1}, & j = s+1. \end{cases}$$

Todistetaan ensin palautuskaava

$$(s-k+1)(f(s+1,j)-f(s,j)) = g(s,j+1) - g(s,j) \tag{39}$$

kaikilla $j \geq k$.

Jos $j = k = 0$, kaavan (39) vasen puoli on

$$(s+1)\left[\binom{-1}{-1}\binom{s}{0} - \binom{-1}{-1}\binom{s+1}{0}\right] = 0,$$

ja sen oikea puoli on $g(s,1)-g(s,0)$. Selvästi $g(s,0) = 0$. Jos $s = 0$, määritelmän perusteella $g(s,1) = 0$. Jos $s \geq 0$, $f(s,1) = 0$, joten $g(s,1) = 0$.

[3]Osoite tätä kirjoitettaessa `http://www.math.temple.edu/%7Ezeilberg/programs.html`

Jos $k \leq j \leq s$ ja $j > 0$, kaavan (39) vasen puoli on

$$(s-k+1)(-1)^{j-k}\binom{j-1}{k-1}\left[\binom{s+1}{j}-\binom{s}{j}\right] = (s-k+1)(-1)^{j-k}\binom{j-1}{k-1}\binom{s}{j-1}.$$

Jos $j = s+1$, vasen puoli on

$$(s-k+1)(-1)^{s-k+1}\binom{s}{k-1}\left[\binom{s+1}{s+1}-\binom{s}{s+1}\right] = (s-k+1)(-1)^{s-k+1}\binom{s}{k-1}.$$

Jos taas $j > s+1$, vasen puoli on 0.

Kaavan (39) oikean puolen laskemiseksi sievennetään ensin lauseketta $g(s,j)$ oletuksilla $k \leq j \leq s$ ja $j > 0$.

$$\begin{aligned}
g(s,j) &= \frac{(-1)^{j-k}(j-k)j}{j-s-1}\cdot\frac{(j-1)!}{(k-1)!(j-k)!}\cdot\frac{s!}{j!(s-j)!}\\
&= \frac{(-1)^{j-k}s!}{(j-s-1)(s-j)!(k-1)!}\cdot\frac{j(j-1)!}{j!}\cdot\frac{j-k}{(j-k)!}\\
&= \frac{(-1)^{j-k+1}(j-k)s!}{(s-j+1)!(k-1)!(j-k)!} = (-1)^{j-k+1}(j-k)\binom{s}{j-1}\binom{j-1}{k-1}.
\end{aligned}$$

Tästä seuraa myös lausekkeelle $g(s,j+1)$ lisäoletuksella $j+1 \leq s$

$$\begin{aligned}
g(s,j+1) &= (-1)^{j-k}(j-k+1)\binom{s}{j}\binom{j}{k-1}\\
&= (-1)^{j-k}(s-j+1)\binom{s}{j-1}\binom{j-1}{k-1},
\end{aligned}$$

missä käytettiin hyväksi binomikerrointen kertomaesityksestä seuraavaa tulosta

$$\binom{s}{j}\binom{j}{k-1} = \frac{s-j+1}{j-k+1}\binom{s}{j-1}\binom{j-1}{k-1}.$$

Siis jos $k \leq j < s$ ja $j > 0$, kaavan (39) oikea puoli on

$$(-1)^{j-k}\binom{s}{j-1}\binom{j-1}{k-1}\big((s-j+1)+(j-k)\big)$$

eli sama kuin vasen puoli.

Jos $j = s$, kaavan (39) vasen puoli on

$$(s-k+1)(-1)^{s-k}\binom{s-1}{k-1}\binom{s}{s-1} = (-1)^{s-k}s(s-k+1)\binom{s-1}{k-1}$$

ja oikea puoli

$$
\begin{aligned}
g(s,s+1)-g(s,s) &= (-1)^{s-k}(s-k+1)\binom{s}{k-1} - (-1)^{s-k+1}(s-k)\binom{s}{s-1}\binom{s-1}{k-1} \\
&= (-1)^{s-k}s\binom{s-1}{k-1} + (-1)^{s-k}(s-k)s\binom{s-1}{k-1} \\
&= (-1)^{s-k}s(s-k+1)\binom{s-1}{k-1}.
\end{aligned}
$$

Jos $j = s+1$, oikea puoli on

$$
g(s,s+2)-g(s,s+1) = (-1)^{s-k+1}(s-k+1)\binom{s}{k-1},
$$

siis sama kuin vasen puoli.

Kun nyt sijoitetaan kaavaan (39) kaikki kokonaisluvut $j \geq \ell+1$ ja lasketaan saadut kaavat puolittain yhteen, vasemman puolen summa on $(s-k+1)(F(s+1)-F(s))$ ja oikean puolen summa on $-g(s,\ell+1)$. Jälkimmäinen summa on tapauksessa $\ell \neq s$

$$
\begin{aligned}
&\frac{-(\ell-k+1)(\ell+1)}{\ell-s}(-1)^{\ell-k+1}\binom{\ell}{k-1}\binom{s}{\ell+1} \\
&\qquad = (-1)^{\ell-k+1}(\ell-k+1)\frac{\ell+1}{s-\ell}\binom{s}{\ell+1}\binom{\ell}{k+1} \\
&\qquad\qquad = (-1)^{\ell-k+1}(\ell-k+1)\binom{s}{\ell}\binom{\ell}{k+1}
\end{aligned}
$$

ja tapauksessa $\ell = s$

$$
(-1)^{s-k+1}(s-k+1)\binom{s}{k-1} = (-1)^{\ell-k+1}(\ell-k+1)\binom{\ell}{k-1}.
$$

Koska $\binom{s}{\ell} = 1$, kun $\ell = s$, väite on todistettu kummassakin tapauksessa. □

Seuraavaa lemmaa 30 tarvitaan lauseen 23 (sivulla 44) todistuksessa. Lemma on peräisin lähteestä [Gra94], jossa se todistetaan perinteisellä kaavojen manipuloinnilla. Seuraavasta todistuksesta suuri osa saadaan seuraavassa lemman 29 todistuksen avulla. Todistus voi näyttää pitkältä, mutta se sisältää vain lähes mekaanista laskujen tarkistamista.

Lemma 30. *Olkoot $m,k,s \in \mathbb{N}$ sellaiset luvut, että $0 < k < m$. Silloin*

$$
(-1)^m \sum_{j=k}^{m-1} (-1)^j \binom{j-1}{k-1}\binom{s}{j} + \sum_{j=m}^{s}\binom{j-1}{k-1}\binom{j-k-1}{m-k-1} = \begin{cases} 0, & \text{jos } s < k, \\ (-1)^{m+k}, & \text{jos } s \geq k. \end{cases}
$$

Todistus. Määritellään

$$G(t) = (-1)^m \sum_{j=k}^{m-1} (-1)^j \binom{j-1}{k-1} \binom{t}{j}.$$

Todistetaan ensin palautuskaava

$$G(t+1) - G(t) = -\binom{t}{k-1} \binom{t-k}{m-k-1}. \tag{40}$$

Tapauksessa $t = k-1$ oikea puoli on $-\binom{-1}{m-k-1} = (-1)^{m-k}$. Silloin vasemmasta puolesta saadaan

$$\begin{aligned} G(t+1) - G(t) &= G(k) - G(k-1) \\ &= (-1)^m \sum_{j=k}^{m-1} (-1)^j \binom{j-1}{k-1} \left[\binom{k}{j} - \binom{k-1}{j} \right] \\ &= (-1)^m \sum_{j=k}^{m-1} (-1)^j \binom{j-1}{k-1} \binom{k-1}{j-1}. \end{aligned}$$

Lauseen 1(f) nojalla $\binom{j-1}{k-1}\binom{k-1}{j-1} = \binom{j-1}{j-1}\binom{0}{k-j}$. Tämä lauseke katoaa, paitsi kun $j = k$, jolloin sen arvo on 1. Siis vasen puolikin on $(-1)^{m-k}$.

Tapauksessa $t \neq k-1$ kirjoitetaan

$$G(t) = (-1)^{m+k} \sum_{j=k}^{m-1} f(t, j),$$

missä f on lemman 29 todistuksessa määritelty funktio. Sille pätee palautuskaava (39) kaikilla $j \geq k$. Kun sijoitetaan tähän kaavaan kaikki kokonaisluvut j, joille $k \leq j \leq m-1$, ja lasketaan saadut yhtälöt puolittain yhteen, saadaan summan vasemman puolen arvoksi $(-1)^{m+k}(t-k+1)(G(t+1) - G(t))$ ja oikean puolen arvoksi $g(t, m) - g(t, k)$. Suoraan g:n määritelmästä nähdään, että $g(t, k) = 0$.

Osoitetaan nyt, että

$$g(t, m) = (-1)^{m+k+1}(t-k+1) \binom{t}{k-1} \binom{t-k}{m-k-1}, \tag{41}$$

kaikilla $t \neq k-1$. Käsitellään kaksi eri tapausta t:n arvon mukaan.

Tapaus 1: $t = m-1$. Tällöin

$$g(t, m) = (-1)^{m-k-1}(t-k+1) \binom{t}{k-1}.$$

Koska $t = m-1$, $\binom{t-k}{m-k-1} = 1$, ja (41) pätee.

Tapaus 2: $t \neq m-1$. Tällöin

$$\begin{aligned}
g(t,m) &= (-1)^{m-k}\frac{(m-k)m}{m-t-1}\binom{m-1}{k-1}\binom{t}{m} \\
&= (-1)^{m-k}\frac{(m-k)(m-k+1)}{m-t-1}\left[\frac{m}{m-k+1}\binom{m-1}{k-1}\right]\binom{t}{m} \\
&= (-1)^{m-k}\frac{(m-k)(m-k+1)}{m-t-1}\binom{m}{k-1}\binom{t}{m} \\
&= (-1)^{m-k}\frac{(m-k)(m-k+1)}{m-t-1}\binom{t}{k-1}\binom{t-k+1}{m-k+1} \\
&= (-1)^{m-k}\frac{(m-k)(t-k+1)}{m-t-1}\binom{t}{k-1}\binom{t-k}{m-k} \\
&= (-1)^{m-k+1}(t-k+1)\binom{t}{k-1}\left[\frac{m-k}{t-m+1}\binom{t-k}{m-k}\right] \\
&= (-1)^{m-k+1}(t-k+1)\binom{t}{k-1}\binom{t-k}{m-k-1}
\end{aligned}$$

Jälleen (41) pätee, ja ja palautuskaava (40) on todistettu kaikilla $t \neq k-1$.

Jos $s < k$, väite on selvästi tosi, koska $G(t) = 0$ kaikilla $t < k$. Jos $s \geq k$, saadaan

$$\begin{aligned}
G(s) = G(s) - G(k-1) &= -\sum_{j=k-1}^{s-1}\binom{j}{k-1}\binom{j-k}{m-k-1} \\
&= -\sum_{j=k}^{s}\binom{j-1}{k-1}\binom{j-k-1}{m-k-1}
\end{aligned}$$

ja siis

$$\begin{aligned}
G(s) &+ \sum_{j=m}^{s}\binom{j-1}{k-1}\binom{j-k-1}{m-k-1} \\
&= -\sum_{j=k}^{m-1}\binom{j-1}{k-1}\binom{j-k-1}{m-k-1} \\
&= -\binom{k-1}{k-1}\binom{k-k-1}{m-k-1} \\
&= (-1)^{m-k}.
\end{aligned}$$

Toiseksi viimeinen yhtäsuuruus seuraa siitä, että kaikilla kyseeseen tulevilla j pätee $j - k - 1 < m-k-1$, ja ainoastaan tapauksessa $j = k$ jälkimmäinen binomikerroin poikkeaa nollasta. $\square$

Viitteet

[AIS93] Rakesh Agrawal, Tomasz Imielinski ja Arun Swami. Mining association rules between sets of items in large databases. Teoksessa *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD '93)*, sivut 207–216, 1993.

[AMS+96] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen ja A. Inkeri Verkamo. Fast discovery of association rules. Teoksessa Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth ja Ramasamy Uthurusamy, (toim.), *Advances in Knowledge Discovery and Data Mining*, luku 12, sivut 307–328. AAAI Press, 1996.

[EN94] Ramez Elmasri ja Shamkant B. Navathe. *Fundamentals of Database Systems.* The Benjamin/Cummings Publishing Company, Inc., toinen painos, 1994.

[GKP90] Ronald L. Graham, Donald E. Knuth ja Oren Patashnik. *Concrete Mathematics.* Addison-Wesley Publishing Company, 1990.

[Gra93] David A. Grable. Sharpened Bonferroni inequalities. *Journal of Combinatorial Theory*, Series B 57:131–137, 1993.

[Gra94] David A. Grable. Hypergraphs and sharpened sieve inequalities. *Discrete Mathematics*, 132:75–82, 1994.

[GS96] Janos Galambos ja Italo Simonelli. *Bonferroni-type Inequalities with Applications.* Probability and its Applications. Springer, 1996.

[Knu73] Donald Knuth. *Fundamental Algorithms*, osa 1 sarjassa *The Art of Computer Programming.* Addison-Wesley Publishing Company, toinen painos, 1973.

[MT96] Heikki Mannila ja Hannu Toivonen. Multiple uses of frequent sets and condensed representations: Extended abstract. Teoksessa Evangelos Simoudis, Jiawei Han ja Usama Fayyad, (toim.), *Second International Conference on Knowledge Discovery and Data Mining*, sivut 189–194. The AAAI Press, August 1996.

[PWZ97] Marko Petkovšek, Herbert S. Wilf ja Doron Zeilberger. *A=B.* A K Peters, Ltd., 1997.

[Wor82] K. J. Worsley. An improved Bonferroni inequality and applications. *Biometrika*, 69(2):297–302, 1982.