
Vastaus asianajotoimisto Risto Kurki-Suoniolta saamaani 29.3. saamaani selvityspyyntöön

6.4.1999

Prof. Ilkka Haikala,
TTKK/Ohjelmistotekniikka
PL 553
33101 Tampere

(03)3652911
ijh@cs.tut.fi

Pyydätte kirjeessänne 29.3.1999 seuraavia lisäselvityksiä ja täsmennyksiä 19.8.1997 laatimaani VF Partnerin ja Systekin maksupääteohjelmien samankaltaisuutta koskevaan selvitykseen.

1. Systekin asiamies on esittänyt kahdeksan muutoskeinoa, joilla Systekin ohjelma saadaan muutettua VF Partnerin ohjelmaksi. Miten kehittämäni samuusmitta suhtautuu tähän väitteeseen.
2. Muuttuisivatko vertailun tulokset, jos käytettävissä olisi täysin riippumattomasti toteutettu maksupääteohjelmisto.
3. Systekin asiamies on väittänyt, että tutkimukseni perusteella voi osoittaa VF Partnerin ohjelman olevan 50% samanlainen kuin Systekin ohjelma.

Esitän seuraavassa ensin kohdassa 1. lyhyen yhteenvedon edellisessä lausunnossani esittämästäni vertailumenetelmästä ja sen jälkeen vastaukseni yo. kysymyksiin kohdissa 2.-4.

1. Yhteenveto vertailussa käytetystä menetelmästä

Vertailtaessa kahta ohjelmaa, joista ohjelmaa A epäillään ohjelman B kopioksi menetellään seuraavasti.

- A1) Ohjelmien muistipaikat muutetaan ns. luurankomuotoon, jolloin ohjelmiin mahdollisesti tehtyjen "kosmeettisten" muutosten vaikutus häviää.
- A2) Jokaiselle tiedoston A muistipaikalle etsitään sitä eniten muistuttava muistipaikka tiedostosta B. Eniten muistuttava muistipaikka on se, jonka kanssa kehittämäni samuusmitta antaa suurimman prosenttiluvun. Prosenttiluvun laskutapa on esitetty yksityiskohdaisesti aikaisemmassa raportissani (luvussa 2). Lyhyesti sanottuna se perustuu tietokoneen raa'an laskentatehon käyttöön siten, että vertailtavista muistipaikoista verrataan keskenään kaikkia mahdollisia osakäskyjonoja. Löydetyt samankaltaisuudet lasketaan yhteen ja samuusmitta on niiden prosenttiosuus koko muistipaikan koosta.
- A3) Löydetyt muistipaikkaparit lajitellaan samuusprosentin mukaan laskevaan suuruusjärjestykseen ja prosenttiluvut piirretään XY-koordinaatistoon pylväsdigrammina, joka antaa havainnollisen kuvan tiedostoista löytyvistä samankaltaisuuksista.

Edellä kuvatun menettelyn ongelma on se, että ohjelmissa esiintyy väistämättä samankaltaisuutta, jota olen kutsunut edellisessä lausunnossani "satunnaiseksi samanlaisuudeksi". Samuusprosentti voi siis olla hyvinkin suuri vaikka kopiointia ei olisikaan tapahtunut. Tämä johtuu mm. seuraavista syistä:

1. TCL-kieli ja laitteen ohjelmointiympäristö ohjaavat ja osin pakottavatkin tietyn tyyppisiin ratkaisuihin.
2. Pyrkimys yhdenmukaiseen käyttöliittymään ja pankkistandardien noudattaminen ohjaavat tiettyihin ratkaisuihin.

-
3. Kun uuden ohjelman tekee ohjelmoija, joka tuntee aikaisemman ohjelman hyvin, hän päätyy suurella todennäköisyydellä samoihin tai samantapaisiin ratkaisuihin. Hänellä voi myös olla hallussaan muistiinpanoja, koulutusmateriaalia yms., joka ohjaa häntä samanlaisiin ratkaisuihin.

Ongelma on siis seuraava: voidaanko vertailussa havaittu samanlaisuus selittää em. seikoilla, vai onko samanlaisuutta niin paljon, että kopioinnin voi perustellusti arvella tapahtuneen. Tämän seikan päättelyyn täytyisi siis oikeastaan olla kolme eri ohjelmaa: ohjelma B ja sen kopioksi väitetty ohjelma A, sekä vertailua varten vielä ohjelma X, josta kiistatta tiedetään, että

- se ei ole kopio ohjelmasta B,
- se on toiminnallisuudeltaan sama kuin ohjelmat A ja B ja
- sen ovat tehneet ohjelmoijat, jotka tuntevat ohjelman B erinomaisesti.

Nyt ohjelmaa B voitaisiin vertailla sekä ohjelmaan A että ohjelmaan X. Kopiointi olisi tällöin osoitettavissa sillä, että samuusmitan antamat samuusprosentit olisivat vertailussa B vs. A merkittävästi suurempia kuin vertailussa B vs. X.

Ongelmaksi muodostuu nyt se, että kiistatta ohjelman X ominaisuuksia täyttävää ohjelmaa ei ole olemassa (edellisessä lausunnossani käytin tosin yhtenä vertailukohtana VF Partnerin uutta ohjelmaa, lausunnon kuvat 8 ja 9).

Eräänlaisena korvikkeena vertailuohjelmalle X päädyin täydentämään menetelmää vielä seuraavasti.

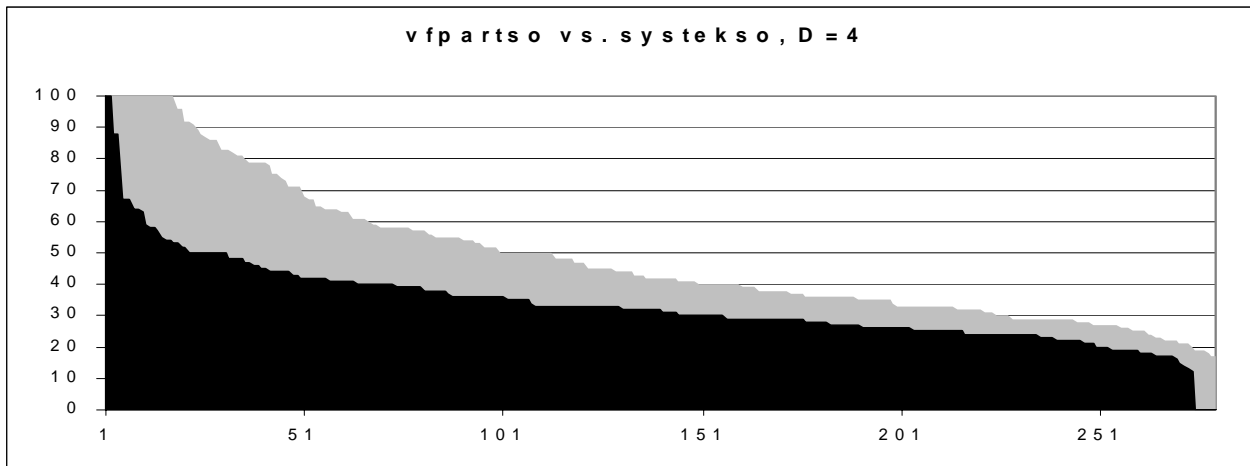
- A4) Etsitään jokaiselle tiedoston A muistipaikalle myös toiseksi eniten sitä muistuttava muistipaikka. Näin saadut samuusprosentit käsitellään samoin kuin edellä askeleessa A3).

Tämä perustuu seuraavaan päättelyyn. Jos eniten muistuttava muistipaikka oletetaan kopioksi, ei sitä toiseksi eniten muistuttava muistipaikka käytännössä voi olla kopio. (Sillä eihän muistipaikka voi olla kahden eri muistipaikan kopio, kuin vain siinä tapauksessa, että samansisältöinen muistipaikka esiintyy ohjelmassa kahdesti. Tämä on erittäin harvinaista.). Vertailukohtana käytetään siis ohjelmaa, josta kopioiksi väitetty kohta on poistettu.

Em. menettelyn ongelma on se, että löydetty eniten muistuttava muistipaikka useissa tapauksissa toteuttaa saman toiminnallisuuden molemmissa ohjelmissa (esimerkiksi sanoman tarkistussumman laskenta). Edellä mainittu vaatimus siitä, että vertailuohjelman X on toteutettava sama toiminnallisuus kuin ohjelmat A ja B ei siis toteudu. Tästä syystä saatu tulos on alaraja satunnaisen samuuden määrälle.

Esimerkki vertailun lopputuloksesta on oheisessa kuvassa 1 (edellisen lausunnon kuva 8). Kuvassa näkyvän harmaan alueen yläreuna kertoo samuusmitan arvot eniten toisiaan muistuttavien muistipaikkojen osalta (koska pylväitä on muutamia satoja, ovat ne kuvassa sulautuneet yhtenäiseksi harmaaksi alueeksi). Vastaavasti mustan alueen yläreuna kuvaa tilanteen toiseksi eniten toisiaan muistuttavien muistipaikkojen osalta. Musta alue kuvaa siis satunnaista samuutta, joka syntyy väistämättä, kun tämäläisiä ohjelmia kirjoitetaan TCL-kielellä. Jossain määrin sen korkeuteen voi vaikuttaa se, että ohjelmoijat kirjoittavat samantapaisissa tilanteissa yhä uudelleen samantapaisia käskysarjoja. Ehkä hieman yllättävä olikin havainto, että kahdessa tapauksessa myös toiseksi eniten muistuttava muistipaikka oli luurankomuodossa täsmälleen sama kuin eniten muistuttava muistipaikka (samuusprosentti 100%).

Harmaan alueen reuna on korkeammalla, koska siihen tulee mukaan myös se samuus, joka aiheutuu siitä, että ohjelmat toteuttavat saman toiminnallisuuden, ja samat ohjelmoijat toteuttavat saman toiminnallisuuden standardeista yms. johtuen usein samalla tai lähes samalla tavalla. Mahdollisen kopioinnin pitäisi näkyä kuvassa mustan ja harmaan alueen suurena erona. Edellisessä lausunnossani esittämieni tutkimuksien ja erilaisten kontrollivertailujen perusteella päädyin tulokseen, jonka mukaan alueiden korkeuserojen perusteella ei voi päätellä VF Partnerin ohjelmaa Systemin ohjelman kopioksi.



Kuva 1: Tiedostojen VFPARTSO ja SYSTEKSO vertailu (edellisen lausunnon kuva 8).

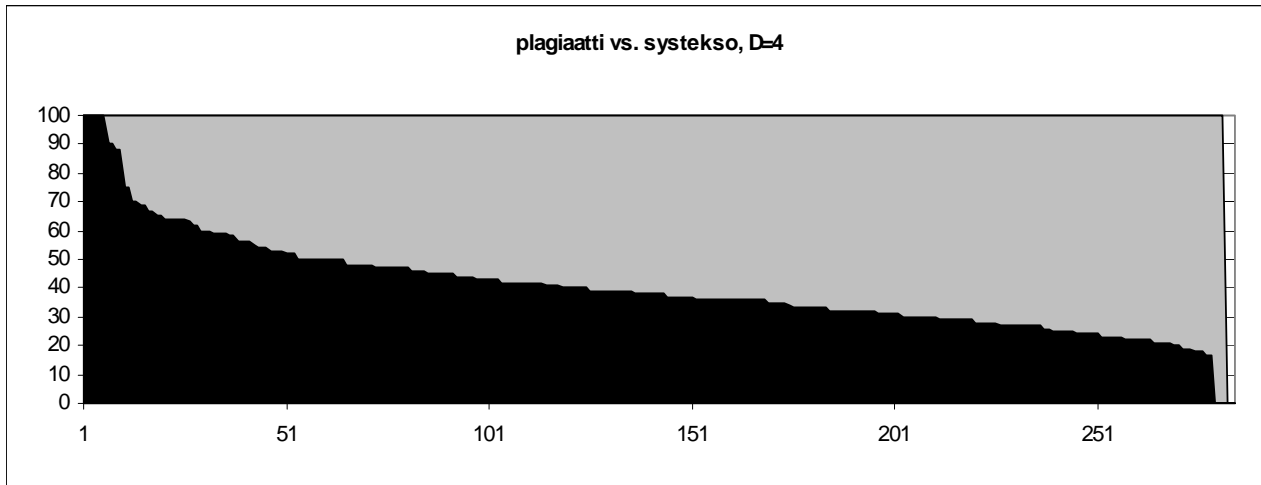
2. Kysymys 1: Voiko samuusmittaa hämätä tekemällä systemaattisia muutoksia

Kaikkien Systemin asiamiehen valituskirjelmässä esittämien muutostapojen osalta vastaus on ehdottomasti ei. Kehitin samuusmitan alunperin nimenomaan tämäntapaisten hämäsyrytysten eliminoimiseksi. Luurankomuodossa koodiin jätetään vain käskyjen operaatiokoodit, kukin omalle omalle rivilleen isoilla kirjaimilla kirjoitettuna. Tällöin esitettyjen muutosten vaikutus eliminoiduu.

Perustelen vastauksen seuraavassa varmuuden vuoksi jokaisen valituskirjelmässä sivulla 11 esitetyn muutostavan osalta erikseen.

- a) Pienet kirjaimet on muutettu pääasiallisesti isoiksi: luurankomuodossa kaikki pienet kirjaimet muutetaan isoiksi.
- b) Jokainen käsky on siirretty omalle rivilleen: luurankomuodossa käskyt sijoitetaan jokainen omalle rivilleen.
- c) Muistipaikkojen numeroita on muutettu: luurankomuodossa muistipaikkojen numerot poistetaan, ts. ne eivät ole vertailussa lainkaan mukana.
- d) Puskurien numerointia on muutettu: luurankomuodossa puskurien numerot poistetaan.
- e) Muuttujien numerointia on muutettu: luurankomuodossa muuttujien numerot poistetaan.
- f) Kommentteja on muutettu: luurankomuodossa kommentit poistetaan.
- g) Kommentteja on poistettu: luurankomuodossa kommentit poistetaan.
- e) Ohjelmalistaukseen on lisätty tabulointeja: luurankomuodossa tabuloinnin poistetaan.

Jos VF Partnerin ohjelma olisi tehty yo. muutoksia tekemällä, olisi VF Partnerin ohjelma luurankomuodossa siis täsmälleen sama kuin lähtökohtana ollut Systemin ohjelma. Jokaiselle muistipaikalle löytyisi samuusmitalla 100% muistipaikkapari alkuperäisestä ohjelmasta (poikkeuksena ohjelmassa olevat muuttamat neljää käskyä lyhyemmät muistipaikat, jotka jäävät vertailun ulkopuolelle). Tässä tilanteessa edellä esitettyä kuvaa 1 vastaava kaavio näyttää aivan erilaiselta (ks. kuva 2) ja kopiointi paljastuisi armotta. Kuvan 2 harmaa alue kattaa lähes koko kuvan osoittaen, että lähes kaikkille muistipaikoille löytyy 100% kopio alkuperäisestä ohjelmasta. (Kuvan oikeassa reunassa oleva pieni aukko johtuu kolmesta muisti-



Kuva 2: Tiedostojen PLAGIAATTI ja SYSTEKSO vertailu muutosten jälkeen.

paikasta, jotka saavat samuusmitan arvon 0%, koska ne lyhyytensä takia putoavat kokonaan vertailun ulkopuolelle. Vertailussa käytetty samuuskriteerin D minimiarvo on tässäkin edellisen lausuntoni tapaan 4.) Toiseksi eniten toisiaan muistuttavat muistipaikat sensijaan näkyvät kuvan 1 tapaan mustana alueena kuvan alareunassa.

3. Kysymys 2: Muuttuisivatko vertailun tulokset, jos käytettävissä olisi täysin riippumattomasti toteutettu maksupääteohjelmisto

Sain sähköpostitse Verifonen maksupäätteen ohjelmistoja Tanskasta (jan.larsen@point.dk) ja Englannista (Ruth_B2@verifone.com, Ruth Burton). Pikaisen perehtymisen perusteella totesin, että saamani materiaali sisälsi ohjelmista useita eri versioita. Valitsin umpimähkään yhden version molemmista postituksista (Tanska: DAN6203, Englanti: TZ5T0303.bd). Muodostin niistä vertailtavat tiedostot Tanska ja Englanti.

Ohjelmien tietoja on koottu taulukkoon 1. Käskyjen lukumäärän perusteella voi päätellä, että ohjelmat ovat suuruudeltaan samaa kokoluokkaa kuin VF Partnerin ja Systekin ohjelmat, englantilaisen version ollessa jonkin verran niitä suurempi ja tanskalaisen pienempi. Ohjelmien ohjelmakoodia en juurikaan ehtinyt lueskelemaan, mutta pintapuolisen tarkastelun perusteella ohjelmat vaikuttivat asiallisilta ja melko hyvin kommentoiduilta.

Minulle esitetyn kysymyksen tarkoituksena lienee ollut ajatus, että ko. ohjelmia voisi ehkä käyttää vertailukohtana (nimellä X edellä kohdassa 1. viitattu ohjelma). Tällaista riippumatonta vertailukohtaa nämäkään ohjelmat eivät kuitenkaan tarjoa seuraavista syistä:

- pankkiyhteys- yms. käytännöissä on huomattavia maakohtaisia eroja ja
- ohjelmien tekijät eivät tunne alkuperäistä ohjelmaa B.

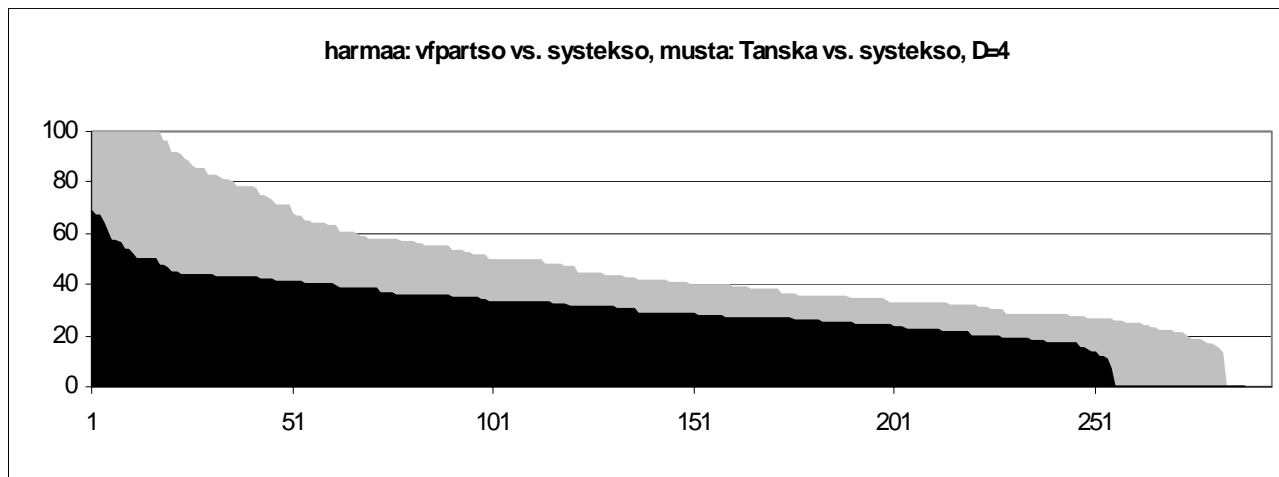
Ohjelmien perusteella saa kuitenkin kuvan siitä, minkä verran samanlaisuutta löytyy ohjelmista, jotka on tehty täysin toisistaan riippumatta, ainoana yhdistävinä tekijöinä TCL-kieli ja samantapainen sovellusalue sekä mahdollisesti samantapainen ohjelmointikoulutus ja/tai koulutusmateriaali.

	SYSTEKSO	VFPARTSO	Tanska	Englanti
Tiedoston koko (tavuina)	152779	175793	369394	530821
Rivejä	6125	7801	11217	18482
Kommentoituja rivejä	2826	3128	9603	13016
Ohjelmakoodimuistipaikkoja	285	333/295 *)	288	391
Käskyjen kokonaismäärä	5871	5812 **)	5245	6972
Komentoimattomat hännät	224 (4%)	586 (8%)	28	109
Kommentoidut hännät	153 (3%)	308 (4%)	171	536

*) osa muistipaikoista esiintyy useaan kertaan

***) VFPARTSO:n samannumeroisten muistipaikkojen käskyt on laskettu vain kertaalleen)

Taulukko 1: Ohjelmien yleisiä ominaisuuksia .



Kuva 3: Tiedostojen Tanska, VFPARTSO ja SYSTEKSO vertailu.

Suoritin muutamia vertailuja, jossa vertasin molempia uusia ohjelmia Systekin ohjelmaan. Esimerkkinä tuloksista on kuva 3. Kuva poikkeaa kuvasta 1 siten, että mustaa aluetta ei ole piirretty toiseksi eniten samankaltaisten muistipaikkojen perusteella, vaan etsimällä jokaiselle tanskalaisen ohjelman muistipaikalle sitä eniten muistuttava muistipaikka ohjelmasta SYSTEKSO. Mustan alueen raja kulkee kuvissa 1 ja 3 suunnilleen samassa kohdassa lähes koko kuvassa. Suurimpia erot ovat käyrien molemmissa päissä. Samuusmitan keskiarvo on mustalla alueella kuvassa 1 31.6% ja kuvassa 3 28.2%. Vertailtaessa englantilaista ohjelmaa tulos oli samansuuntainen. Tulos on mielestäni odotettu, kun otetaan huomioon, että maakohtaisissa käytännöissä on huomattavia eroja, ja että ohjelman tehneillä ohjelmoijilla ei ole mitään yhteyttä Systekin ohjelmaan.

Vastaus kysymykseen on, että nyt käytössäni olleista uusista ohjelmista saamani tulokset eivät muuta aikaisemmassa lausunnossa esittämäni johtopäätöstä. Uusi vertailu näyttää lähinnä vahvistavan uskoa siihen, että käytetty vertailutekniikka on järkevä.

4. Kysymys 3: Voiko tutkimukseni perusteella osoittaa VF Partnerin ohjelman olevan 50% samanlainen kuin Systekin ohjelma.

Vastaus kysymykseen on ei, tai ainakin väite on erittäin harhaanjohtava.

1. Kuten jo edellä totesin, syntyy ohjelmiin monista eri syistä väistämättä samankaltaisuutta jopa tilanteessa, jossa ohjelmoijat eivät ole mitenkään tietoisia toistensa työstä ja sovelletut maa-kohtaiset standardit poikkeavat toisistaan (vrt. kohta 3 edellä, erityisesti kuvan 3 mustan alueen yläreuna).
2. Jos sovellusalueen standardit ovat samat, on samankaltaisuutta syytä odottaa enemmän kuin kuvan 3 mustalla alueella. Jos lisäksi vielä samat ohjelmoijat tekevät saman ohjelman uudelleen tai ainakin tuntevat aikaisemman ohjelman koodin hyvin, tällainen satunnainen samankaltaisuus lisääntyy pakostakin entisestään, vaikka kyseessä ei olisikaan alkuperäisen ohjelman suora kopiointi. Tällöin samuusmitan voi odottaa antavan suurehkojakin positiivisia arvoja (harmaan alueen yläreuna kuvassa 3) ilman, että se on osoitus ohjelmiston kopioinnista.

Verrattaessa VF Partnerin ja Systekin ohjelmia toisiinsa samuusmittojen keskiarvo on n. 46.5%. Esimerkiksi kuvasta 1 voi nähdä, että ohjelmissa on joukko muistipaikkoja, jotka ovat luurankomuodossa vertailtaessa aivan samat (samuus 100%, 18kpl). Seuraavissa n. 80 muistipaikassa samuutta on yli 50% ja loppuissa n. 200 muistipaikassa vähemmän kuin 50%. Orjallinen kopiointi ja systemaattinen muutosten tekeminen hämäysmielessä edellä kohdassa 2. esitetyillä tavoilla tulisi siis kysymykseen vain 18 muistipaikan kohdalla.

Edellä kuvattu samuus ei siis ole ohjelmassa yhtenäisellä alueella siten, että ohjelman 5000 rivistä voisi osoittaa jonkin yhtenäisen alueen/alueita, joilta identtiset käskysarjat löytyvät.