

**METODA *GENERAL ADDITIVE DATA PERTURBATION*
(GADP) UNTUK KEAMANAN DATABASE**

LAPORAN TUGAS AKHIR

Diajukan untuk memenuhi salah satu tugas mata kuliah
EC-7010 Keamanan Sistem Lanjut

Oleh :

Mochamad Widiyanto

NIM : 23203 117



**INSTITUT TEKNOLOGI BANDUNG
2004**

ABSTRAK

METODA *GENERAL ADDITIVE DATA PERTURBATION* (GADP) UNTUK KEAMANAN DATABASE

Oleh :

Mochamad Widiyanto

**Departemen Teknik Elektro
Institut Teknologi Bandung**

Keamanan database suatu organisasi telah mendapatkan perhatian serius selama beberapa tahun terakhir. Hal ini didorong oleh beberapa hal, seperti adanya peningkatan jumlah data yang disimpan dalam database, perkembangan teknik-teknik analisis data serta kebutuhan akan keamanan data yang disimpan. Metoda *perturbation* data biasanya digunakan untuk memproteksi data *konfidensial* yang bersifat numerik dari *query* tidak sah, sementara harus membuka akses secara maksimum dan menyediakan informasi secara akurat yang dibutuhkan oleh *query* dengan otorisasi legal.

Untuk menyediakan informasi secara akurat, harus dipertimbangkan bahwa *perturbation* yang digunakan tidak akan mengubah nilai relasi antar atribut. Dengan adanya atribut non-konfidensial dalam database, beberapa metoda *perturbation* menyebabkan terjadinya perubahan nilai relasi tersebut. Metoda *General Additive Data Perturbation* (GADP) merupakan metoda *perturbation* yang tidak akan menyebabkan perubahan nilai relasi antar atribut dalam database. Metoda-metoda *perturbation* yang tergolong *Additive Data Perturbation* (ADP) dapat ditunjukkan sebagai kasus spesial dari metoda GADP.

Pada database yang memiliki distribusi normal *multivariate*, metoda GADP dapat menyediakan tingkat keamanan maksimum dengan bias minimum.

Kata kunci : manajemen database, keamanan data, *perturbation* data.

KATA PENGANTAR

Segala puji milik Allah swt yang telah mencurahkan nikmat yang tak terhingga kepada setiap hamba-Nya, baik yang diminta maupun yang tak diminta. Karena pertolongan-Nya semata maka Laporan Tugas Akhir "*Metoda General Additive Data Perturbation Untuk Keamanan Database*" ini dapat terwujud. Salawat dan salam semoga tercurah kepada Nabi Muhammad saw yang telah mendidik umatnya untuk memahami dan melaksanakan aspek lahir dan batin dari setiap ritual ibadah.

Pada kesempatan ini kami sampaikan terima kasih yang sebesar-besarnya kepada semua pihak yang telah membantu penyelesaian laporan tugas akhir ini. Terima kasih kami sampaikan kepada :

1. Bapak Dr. Ir. Budi Rahardjo, sebagai dosen mata kuliah EC-7010 Keamanan Sistem Lanjut, yang telah memberikan kuliah dan bimbingan hingga terselesaikannya tugas ini.
2. Rekan-rekan sejawat sesama mahasiswa S2 TI Dikmenjur yang telah membantu penyusunan laporan tugas akhir ini.
3. Semua pihak yang tidak dapat kami sebutkan satu-persatu, yang telah membantu penyelesaian laporan tugas akhir ini.

Semoga segala dukungan yang diberikan selalu mendapatkan balasan yang baik dari Allah swt.

Kekurangan-kekurangan dan ketidak-sempurnaan tentu masih melekat dalam laporan ini. Berkaitan dengan hal tersebut maka dengan penuh kerendahan hati kami mohon saran, kritik dan masukan-masukan dalam rangka memperbaiki laporan ini.

Kami berharap semoga laporan tugas akhir ini dapat memberikan kontribusi dan membawa manfaat bagi kita semua.

Bandung, 27 Desember 2004
Penyusun,

DAFTAR ISI

	Halaman
ABSTRAK	i
KATA PENGANTAR	ii
DAFTAR ISI	iii
DAFTAR TABEL	v
DAFTAR SINGKATAN	vi
DAFTAR NOTASI	vii
BAB I PENDAHULUAN	1
BAB II <i>BIAS DAN KEAMANAN</i>	5
2.1. Pengukuran <i>Bias</i> Yang Diakibatkan Oleh <i>Perturbation</i>	5
2.2 Pengukuran Keamanan Yang Dapat Disediakan Oleh <i>Perturbation</i>	6
2.2.1. Pengukuran Keamanan Untuk Sebuah Atribut	7
2.2.2. Pengukuran Keamanan Untuk Kombinasi <i>Linear</i>	7
BAB III PERFORMA METODA-METODA DATA <i>PERTURBATION</i>	11
3.1. Menentukan <i>Bias</i> Dan Persyaratan Keamanan Untuk <i>Perturbation</i>	11
3.2. Metoda <i>Additive Data Perturbation</i>	12
3.2.1. Metoda <i>Simple Additive Data Perturbation</i> (SADP)	12
3.2.2. Metoda <i>Correlated-Noise Additive Data Perturbation</i> (CADP)	14
3.2.3. Metoda <i>Bias-Corrected Correlated-Noise Additive Data Perturbation</i> (BCADP)	15
3.3. <i>Multiplicative Data Perturbation</i> (MDP)	17

BAB IV METODA <i>GENERAL</i> UNTUK <i>ADDITIVE DATA PERTURBATION</i>	19
4.1. Kebutuhan Akan Metoda <i>Perturbation</i> Baru	19
4.2. Metoda <i>General Additive Data Perturbation</i> (GADP)	20
4.3. Landasan Metoda GADP	21
4.4. Metoda GADP Sebagai Bentuk Umum Dari Metoda <i>Additive Perturbation</i>	21
4.5. Fleksibilitas Metoda GADP	22
BAB V PERFORMA METODA GADP	24
5.1. Implementasi Metoda GADP	24
5.2. Performa <i>Bias</i> Dan Keamanan Pada Metoda GADP	27
5.2.1. <i>Bias</i>	27
5.2.2. Evaluasi Tingkat Keamanan	28
5.3. Performa Metoda GADP Pada Populasi Berdistribusi <i>Log-Normal</i>	29
BAB VI KESIMPULAN	30
DAFTAR PUSTAKA	32

DAFTAR TABEL

	Halaman
Tabel 1. <i>Database bank</i>	2
Tabel 2. <i>Pengukuran deskriptif database bank.</i>	3
Tabel 3. <i>Karakteristik statistik dari berbagai metoda perturbation.</i>	12
Tabel 4. <i>Pengukuran deskriptif database bank setelah dilakukan metoda perturbation SADP.</i>	13
Tabel 5. <i>Pengukuran deskriptif database bank setelah dilakukan metoda perturbation CADP.</i>	15
Tabel 6. <i>Pengukuran deskriptif database bank setelah dilakukan metoda perturbation BCADP.</i>	16
Tabel 7. <i>Pengukuran deskriptif database bank setelah dilakukan metoda perturbation MDP.</i>	18
Tabel 8. <i>Pengukuran deskriptif database bank setelah dilaksanakan metoda perturbation GADP</i>	28

DAFTAR SINGKATAN

Singkatan Yang Digunakan

ADP	<i>Additive data perturbation</i>
BCADP	<i>Bias-corrected correlated-noise additive data perturbation</i>
CADP	<i>Correlated-noise additive data perturbation</i>
DBA	<i>Database administrator</i>
GADP	<i>General additive data perturbation</i>
MDP	<i>Multiplicative data perturbation</i>
SADP	<i>Simple additive data perturbation</i>
SDB	<i>Statistical database system</i>

DAFTAR NOTASI

Notasi Yang Digunakan

X	merepresentasikan atribut konfidensial <i>original</i>
S	merepresentasikan atribut non-konfidensial
Y	merepresentasikan atribut konfidensial <i>yang telah diberi perturbation</i>
D	merepresentasikan matriks diagonal dimana elemen-elemen diagonalnya adalah sama dengan yang ada pada X
e	merepresentasikan <i>noise</i>
d	merepresentasikan <i>level</i> dari perturbation
d_1	$= (1 + d)^{0,5}$
d_2	$= d_1 - 1$
\sum_{XX}	merepresentasikan matriks <i>covariance</i> dari X
\sum_{SS}	merepresentasikan matriks <i>covariance</i> dari S
\sum_{YY}	merepresentasikan matriks <i>covariance</i> dari Y
\sum_{ee}	merepresentasikan matriks <i>covariance</i> dari e
\sum_{XS}	merepresentasikan matriks <i>covariance</i> dari (X dan S)
\sum_{XY}	merepresentasikan matriks <i>covariance</i> dari (X dan Y)
\sum_{YS}	merepresentasikan matriks <i>covariance</i> dari (Y dan S)

BAB I

PENDAHULUAN

Database memainkan peranan yang sangat penting dalam sebuah organisasi masa kini dengan memberikan dukungan untuk penyimpanan data esensial dan untuk proses pengambilan keputusan. Menyediakan kenyamanan akses ke database untuk *user* yang memiliki otorisasi merupakan tujuan penting yang seharusnya dapat dicapai oleh administrator database (*database administrator/ DBA*). Database organisasi berisi beragam data, baik data yang bersifat konfidensial maupun yang bersifat non-konfidensial. Dilema besar yang dihadapi oleh DBA adalah bagaimana melindungi data yang bersifat konfidensial dan sensitif tanpa menghalangi akses ke data yang dibutuhkan untuk proses pengambilan keputusan. Dilema ini kini menjadi semakin penting dengan adanya keharusan untuk menjaga privasi data yang menyangkut pribadi seseorang.

Salah satu cara untuk memecahkan masalah ini adalah dengan mempergunakan sistem database statistik (*statistical database system/ SDB*). Dalam suatu SDB, *user* diijinkan untuk memperoleh hanya angka-angka statistik dari suatu subset data. Jika suatu database hanya berisi data yang bersifat konfidensial, maka pendekatan yang umum dilakukan adalah dengan membatasi *user* tertentu untuk memperoleh angka-angka statistik dan mengijinkan *user* yang lain atau program aplikasi untuk mengakses data secara individual.

Sebagai ilustrasi, diambil contoh kasus database sebuah bank. Diasumsikan bahwa database bank berisi data tentang nasabah dan tabungan yang dimilikinya dalam bank tersebut (saldo, debit dan kredit). Atribut-atribut ini merupakan data non-konfidensial (bagi karyawan bank tersebut). Selain itu, bank juga mengumpulkan informasi lain, seperti *home equity*, investasi pada saham/ obligasi serta *liabilities* dari nasabah, dengan jaminan eksplisit bahwa

informasi tersebut hanya akan dipergunakan secara kolektif (berbentuk angka-angka statistik) dan data individual nasabah tersebut tidak akan diberitahukan kepada siapapun, termasuk pegawai bank sendiri. Atribut-atribut ini dengan demikian menjadi data yang bersifat konfidensial. Bank dapat mempergunakan informasi dalam database ini untuk mempersiapkan layanan baru atau meningkatkan layanan tertentu bagi para nasabahnya.

Tabel 1. *Database bank.*

Nasabah	<i>Home Equity</i>	Saham/ Obligasi	<i>Liabilities</i>	Simpanan	Kredit
1	120,62	48,86	73,72	22,29	37,75
2	120,04	54,21	88,65	18,63	53,69
	-	-		-	
	-	-		-	
9.999	86,96	50,01	86,98	16,52	44,09
10.000	81,92	50,41	67,73	21,04	51,48

Sebagian database bank tersebut diperlihatkan pada Tabel 1. Diasumsikan bahwa database ini memiliki distribusi normal *multivariate*. Deskripsi statistik dari atribut-atribut pada database ini ditunjukkan pada Tabel 2. Karena tidak mungkin untuk menggambarkan setiap tipe *query* yang mungkin diminta oleh *user*, maka *summary* ini mengukur representasi respons terhadap *query* yang dianggap penting untuk proses pembuatan keputusan.

Dilema yang dihadapi oleh DBA pada ilustrasi ini cukup jelas. Bank membutuhkan data ini untuk mengambil keputusan, maka penting untuk menyediakan respons yang akurat untuk setiap *query* yang dilakukan oleh pegawainya mengenai data secara kolektif dan angka-angka statistiknya. Dilain pihak, bank harus tetap menjaga privasi individual nasabahnya.

Organisasi memiliki *policy* dan prosedur untuk menjaga akses tidak sah, baik yang dilakukan oleh “*orang dalam*” maupun oleh “*orang luar*”. Akses tidak sah yang dilakukan oleh orang dalam adalah lebih berbahaya, mengingat

bahwa lebih dari 75% *computer abusement* dilakukan oleh orang dalam sendiri. Fokus dari paper ini adalah menjaga akses tidak sah yang dilakukan oleh orang dalam. Digunakan istilah “*snooper*” untuk entitas yang memiliki ijin/ otoritas yang sah untuk mengakses database namun menggunakan akses sah ini untuk mengumpulkan data individual dengan atribut konfidensial yang seharusnya tidak boleh untuk diketahui.

Tabel 2. Pengukuran deskriptif database bank.

Summary Statistik		Atribut	Matriks <i>Covariance</i> (Korelasi ditunjukkan dalam kurung)				
Mean (Rp.000)	Std. Dev. (Rp. 000)		Home Equity	Saham/ Obligasi	Liabilities	Simpanan	Kredit
100,00	20,00	Home Equity	400,00 (1,00)				
50,00	10,00	Saham/ Obligasi	140,00 (0,70)	100,00 (1,00)			
80,00	20,00	Liabilities	320,00 (0,80)	150,00 (0,75)	400,00 (1,00)		
20,00	5,00	Simpanan	50,00 (0,50)	20,00 (0,40)	25,00 (0,25)	25,00 (1,00)	
50,00	10,00	Kredit	60,00 (0,30)	20,00 (0,20)	30,00 (0,15)	30,00 (0,60)	100,00 (1,00)

Teknik kontrol akses sering diimplementasikan untuk menyediakan *summary* statistik data tanpa membuka data individualnya sendiri. Walaupun teknik ini dapat menjaga terbukanya data konfidensial dari beberapa tipe *query*, namun hal ini tidak menjamin terbukanya data konfidensial dari semua tipe *query*. *Snooper* melalui penggunaan *query* yang sah dan kemampuan analisis statistik yang dimilikinya, mungkin dapat mengidentifikasi, melalui entitas yang diberikan, nilai yang tepat dari suatu atribut konfidensial. Karena terbukanya rahasia data konfidensial terjadi melalui kesimpulan (*inference*), maka hal ini dikenal sebagai *inferential disclosure*. Selanjutnya, jika nilai sesungguhnya dari atribut konfidensial telah disimpulkan, maka ini dikenal sebagai *complete disclosure*. Pada contoh database bank diatas, dengan *inferential*, *complete disclosure* akan terjadi jika *snooper* dapat

mengidentifikasi bahwa nasabah #2 memiliki Rp.120.040,- pada kolom *home equity* (lihat Tabel 1).

Metoda *data perturbation* sangat berguna pada kasus dimana data berbentuk numerik dan *complete disclosure* tidak diijinkan. *Data perturbation* merupakan modifikasi terhadap data dalam atribut konfidensial dengan menggunakan *random noise*. Akses *user* dibatasi hanya pada data yang telah dimodifikasi, sehingga akan menjamin nilai asli dari data konfidensial tidak akan diketahui. Tujuan dari *data perturbation* adalah tetap menjaga kerahasiaan data konfidensial ketika akses maksimal terhadap informasi dibutuhkan secara akurat (membatasi *bias*).

Metoda *perturbation* dapat mencegah terjadinya *complete disclosure*, namun metoda ini mungkin masih dapat mengalami *partial disclosure*. *Partial disclosure* adalah kondisi dimana *snooper* dapat memperoleh nilai estimasi dari data beratribut konfidensial yang mendekati nilai data sesungguhnya/ aslinya. *Partial disclosure* dikatakan telah terjadi jika nilai estimasi yang dibuat oleh *snooper* lebih mendekati kenyataan daripada level yang diijinkan oleh DBA.

Problem lain dari metoda *data perturbation* adalah respons untuk *query* yang menggunakan *data perturbation* dapat berbeda dengan respons yang menggunakan data *original*. Hal ini dikenal sebagai *bias* yang diakibatkan oleh *perturbation* (secara singkat akan disebut sebagai *bias*). Dengan kata lain *perturbation* akan mengubah satu atau lebih nilai statistik yang tercantum pada Tabel 2. Secara umum, semakin tinggi level *perturbation* akan semakin tinggi pula *bias* yang terjadi.

Pada paper ini akan dibahas metoda *General Additive Data Perturbation* (GADP) yang dapat menjaga korelasi antar semua atribut (baik atribut konfidensial maupun atribut non-konfidensial) secara sama, baik sebelum maupun sesudah *perturbation*.

BAB II

BIAS DAN KEAMANAN

Pertama akan dibahas mengenai pengukuran *bias* dan *security* (keamanan). Pengukuran kedua hal ini dibutuhkan untuk mengevaluasi pengaruh atribut konfidensial *non-close* pada metoda *perturbation*.

Untuk mengevaluasi metoda *perturbation* apapun, penting untuk mengukur *bias* dan keamanan yang dapat disediakan oleh metoda tersebut. Berikut akan didefinisikan pengukuran *bias* dan keamanan.

2.1. Pengukuran *Bias* Yang Diakibatkan Oleh *Perturbation*

Metoda *perturbation* dilakukan dengan menambahkan *noise* pada nilai *original* untuk membentuk nilai *perturbation*. *Noise* biasanya memiliki *mean* 0 dan *variance* spesifik. Hal ini mengakibatkan terjadinya *bias* pada *variance* dan karakteristik lain, tetapi tidak pada *mean*. Disini diasumsikan bahwa *perturbation* telah mengubah *standard deviasi* dari atribut *home equity* dan *liabilities* dari Rp. 20.000,- ke Rp. 25.000,-, dan saham/ obligasi dari Rp. 10.000,- ke Rp. 12.500,-. Semua aspek lain dari database dibiarkan tidak mengalami perubahan. Uraian berikut akan membahas klasifikasi berbagai tipe *bias*.

Bias tipe A. *Perturbation* yang dilakukan pada atribut tertentu pada database meningkatkan *variance* dari atribut tersebut. Respons dari *query* “Berapa standard deviasi dari *home equity* ?”, dari data yang telah diberi *perturbation* akan dijawab Rp. 25.000,- yang berbeda dari data *original* Rp. 20.000,-. Hal ini juga akan berpengaruh pada *query* yang lain seperti “Berapa 5% *home equity* ?”, “Berapa rata-rata *home equity* dari 5% nasabah paling atas ?” dan sebagainya. *Bias* seperti ini dikenal sebagai *bias* tipe A.

Bias tipe B. *Perturbation* dapat mengubah hubungan antar atribut-atribut konfidensial. *Bias* tipe ini dapat terjadi pada perubahan *variance* dari atribut

konfidensial dan/ atau pada perubahan *covariance* diantara atribut-atribut. Pada contoh diatas, *perturbation* akan mengubah korelasi antara atribut konfidensial *home equity* dan saham/ obligasi dari 0,8 pada database *original* (Tabel 2) ke 0,45. *Bias* seperti ini dikenal sebagai *bias* tipe B.

Bias tipe C. *Perturbation* yang dilakukan pada atribut konfidensial dapat juga mengubah hubungan antara atribut konfidensial dan atribut non-konfidensial. *Bias* ini dapat terjadi pada perubahan *variance* atribut konfidensial dan/ atau pada perubahan *covariance* antara atribut konfidensial dan atribut non-konfidensial. Pada contoh diatas, korelasi antara atribut konfidensial *home equity* dan atribut non-konfidensial *saving/checking*, berubah dari 0,50 pada database *original* ke 0,40 setelah dilakukan *perturbation*. *Bias* ini disebut *bias* tipe C.

Bias tipe D. Jika sebaran database tidak berupa distribusi normal *multivariate* dan/ atau tambahan *noise* tidak normal *multivariate*, bentuk dari database yang telah diberi *perturbation* tidak selalu dapat ditentukan. Jika hal ini terjadi, maka respons terhadap *query* yang melibatkan persentase, jumlah, *conditional mean* dan lain-lain mungkin akan mengalami *bias*. Kondisi ini disebut sebagai *bias* tipe D.

Berikut akan dilakukan pembahasan tentang pengukuran keamanan. Notasi yang digunakan dalam pembahasan dapat dilihat pada Daftar Notasi di halaman vii.

2.2. Pengukuran Keamanan Yang Dapat Disediakan Oleh *Perturbation*

Sebagaimana telah diindikasikan sebelumnya, metoda *perturbation* menjamin tidak akan terjadi *complete disclosure*, namun mungkin masih dapat terjadi *partial disclosure*. Oleh karenanya perlu untuk mengukur tingkat keamanan yang dapat disediakan oleh sebuah teknik *perturbation* terhadap terjadinya *partial disclosure*.

2.2.1. Pengukuran Keamanan Untuk Sebuah Atribut

Secara tradisional, keamanan yang dapat disediakan oleh teknik *perturbation* diukur sebagai *variance* dari perbedaan antara nilai *original* dan nilai yang telah diberi *perturbation*. Ukuran ini diberikan dengan $Var(X - Y)$, dimana X merupakan sebuah atribut *original* dan Y merupakan atribut yang telah diberi *perturbation*. Ukuran ini dapat dibuat dengan menggunakan skala *invariant* yang mengacu pada *variance* X dengan menyatakan keamanan sebagai :

$$S_1 = Var(X - Y) / Var(X) \dots\dots\dots (1)$$

Ukuran ini cocok digunakan untuk mengukur keamanan yang disediakan untuk sebuah atribut terhadap *snooper* yang memiliki akses sangat terbatas ke database. *User* sedemikian (biasanya pegawai setingkat *entry-level* atau *low-level* dengan tanggung jawab yang sangat terbatas dalam pembuatan keputusan) akan mengestimasi nilai data sesungguhnya hanya dengan mempergunakan nilai-nilai yang telah diberi *perturbation*. Karenanya, keamanan yang disediakan dapat diukur sebagai S_1 .

2.2.2. Pengukuran Keamanan Untuk Kombinasi *Linear*

Dalam suatu organisasi terdapat juga *user* yang memiliki otoritas untuk mengakses database secara bebas. *User* demikian biasanya adalah manajer *high-level* yang dapat memperoleh akses tanpa batas ke database (kecuali nilai *original* dari atribut yang bersifat konfidensial). Mereka biasanya bahkan mempunyai pengetahuan tentang atribut-atribut mana yang telah diberi *perturbation* dan seberapa besar *perturbation* yang diberikan. *User* demikian biasanya juga memiliki otoritas untuk mengeksplorasi hubungan antar atribut. *User* dengan level akses yang tinggi dan memiliki pengetahuan sedemikian, jika menjadi *snooper* akan mencoba untuk mengestimasi nilai sesungguhnya dari data yang bersifat konfidensial dengan menggunakan cara-

cara yang lebih *advance* daripada *user* dengan akses yang terbatas. Salah satu cara estimasi yang lebih *advance* adalah dengan mempergunakan kombinasi *linear* dari data yang terdapat pada atribut-atribut yang bersifat konfidensial.

Dalam contoh database bank diatas, terdapat kombinasi *linear* sebagai berikut :

Home Equity + Saham/ Obligasi – Liabilities.

Ini merupakan salah satu contoh kombinasi *linear* dari atribut-atribut konfidensial yang terdapat dalam database bank, yang biasanya dikenal sebagai “investasi *netto* diluar bank”. *Snooper* dapat memperkirakan jumlah kombinasi *linear* ini dengan mempergunakan nilai atribut-atribut konfidensial yang telah diberi *perturbation* maupun atribut-atribut non-konfidensial.

Jika atribut non-konfidensial yang terdapat dalam database dipergunakan untuk mengestimasi kombinasi *linear* dari atribut konfidensial, maka resiko *disclosure* menjadi semakin tinggi. Dalam contoh kasus database bank diatas, *snooper* mungkin mencoba untuk mengestimasi “investasi *netto* diluar bank” ini dengan mempergunakan semua informasi yang dapat diperolehnya, yaitu nilai-nilai yang telah diberi *perturbation* dari atribut konfidensial *home equity*, saham/ obligasi dan *liabilities* serta nilai-nilai atribut non-konfidensial seperti saldo, debit dan kredit. Pada kasus seperti ini, *snooper* mungkin dapat memahami dengan baik *variability* dari kombinasi *linear* pada atribut-atribut konfidensial tersebut, sehingga tingkat keamanan database menjadi berkurang.

Penting untuk dicatat bahwa kombinasi *linear* dari atribut-atribut tidak akan disimpan secara terpisah dalam database karena atribut-atribut ini dihitung mempergunakan atribut lainnya. Dalam kasus seperti ini, DBA mungkin tidak menyadari bahwa kombinasi *linear* boleh jadi akan menyebabkan terjadinya *partial disclosure*. Karena *disclosure* akan menyebabkan menurunnya tingkat keamanan, maka penting untuk mengukur tingkat keamanan dari aspek kombinasi *linear* pada atribut-atribut yang terdapat dalam database.

Pengukuran keamanan terhadap kombinasi *linear* merupakan hal yang sulit, karena dalam database sesungguhnya, terdapat tak terhingga banyaknya kombinasi sedemikian. Pada database bank, selain kombinasi *linear* “investasi *netto* diluar bank” seperti yang telah diuraikan diatas, mudah untuk melihat adanya kombinasi *linear* lainnya. Secara praktis, tidak mungkin untuk mengidentifikasi dan mengukur tingkat keamanan dari semua kombinasi *linear* yang ada dalam suatu database. Pada studi ini akan dipergunakan analisis korelasi *canonical* sebagai pengukuran tingkat keamanan umum terhadap kombinasi *linear* dalam suatu database.

Analisis korelasi *canonical* dapat dipergunakan untuk mengukur proporsi maksimum *variance*, dimana *snooper* dapat menggunakannya untuk setiap kombinasi *linear* dari atribut konfidensial yang tidak diketahui dengan menggunakan kombinasi *linear* dari atribut yang telah diketahui (baik yang berupa atribut non-konfidensial dan/ atau atribut konfidensial yang telah diberi *perturbation*). Karenanya, tingkat keamanan dapat diukur oleh administrator database dengan menggunakan analisis ini. Tingkat keamanan adalah proporsi dari *variance* yang harus ada (tidak dapat dihilangkan) yang akan dipergunakan oleh *snooper* dalam mengestimasi kombinasi *linear* dari atribut-atribut konfidensial. Fitur penting dari pengukuran tingkat keamanan menggunakan analisis ini adalah bahwa penghitungan dapat dilakukan secara langsung untuk database apapun dan metoda *perturbation* apapun.

Diberikan λ sebagai representasi dari *eigenvalue* terbesar yang dihasilkan oleh matriks berikut :

$$\sum_{xx}^{-1} \sum_{xv} \sum_{vv}^{-1} \sum_{vx} \quad \text{dimana } V = \{S, Y\} \dots\dots\dots (2)$$

dan

$$\sum_{vv} = \begin{bmatrix} \sum_{ss} & \sum_{sy} \\ \sum_{ys} & \sum_{yy} \end{bmatrix}$$

Nilai λ menunjukkan proporsi maksimum dari variabilitas setiap kombinasi *linear* dari X yang dapat dijelaskan menggunakan kombinasi *linear* manapun dari Y dan S . Karenanya keamanan terhadap kombinasi *linear* dapat didefinisikan sebagai :

$$S_2 = 1 - \lambda \dots\dots\dots (3)$$

Sehingga, untuk setiap kombinasi *linear* dari X , paling sedikit S_2 proporsi variabilitas tetap tidak dapat dijelaskan. Harus dicatat bahwa rumusan diatas dapat dipergunakan untuk database apapun, untuk metoda *perturbation* apapun dan untuk *user* manapun, sehingga rumus ini merupakan alat yang *powerful* untuk mengevaluasi keamanan.

Akhirnya penting untuk dicatat bahwa dalam konteks organisasi, semua pengukuran keamanan (S_1 dan S_2) akan memungkinkan DBA memiliki informasi penting mengenai efektivitas metoda *perturbation*.

BAB III

PERFORMA METODA-METODA DATA *PERTURBATION*

Pada bagian ini akan diuraikan secara umum perbedaan berbagai metoda *data perturbation* yang ada.

3.1. Menentukan *Bias* Dan Persyaratan Keamanan Untuk *Perturbation*

Untuk memahami performa *bias* dan keamanan dari setiap metoda, diasumsikan bahwa DBA telah menentukan persyaratan keamanan ketika memberikan *perturbation* pada database bank seperti berikut :

- (i) Tidak terjadi *bias* pada *summary* pengukuran atribut-atribut konfidensial individual yang berhubungan dengan perubahan *variance* (tidak terjadi *bias* tipe A).
- (ii) Hubungan antar atribut konfidensial harus tetap sama, baik sebelum maupun setelah *perturbation* (tidak terjadi *bias* tipe B).
- (iii) Hubungan antara variabel konfidensial dan variabel non-konfidensial harus tetap sama, baik sebelum maupun setelah *perturbation* (tidak terjadi *bias* tipe C).
- (iv) Distribusi dari atribut-atribut konfidensial harus tetap sama, baik sebelum maupun setelah *perturbation* (tidak terjadi *bias* tipe D).
- (v) Tingkat keamanan untuk sebuah atribut (atribut tunggal) paling kecil adalah 1,00 untuk setiap atribut.
- (vi) Tingkat keamanan terhadap kombinasi *linear* paling kecil adalah 0,50.

Pada bagian berikut akan digunakan database bank sebagai ilustrasi numerik pada aplikasi berbagai metoda *perturbation*. Walaupun demikian, hasilnya tetap dapat digeneralisir untuk database dengan distribusi normal *multivariate* lainnya.

3.2. Metoda Additive Data Perturbation (ADP)

Metoda ADP dikembangkan dari metoda yang semula digunakan untuk atribut tunggal menjadi metoda *perturbation* untuk *multi-attribut*. Pada jenis kasus apapun, metoda ADP digunakan dengan menambahkan *noise* yang memiliki *mean* 0, sehingga tidak akan menyebabkan terjadinya *bias* pada estimasi *mean*. *Summary* karakteristik yang dihasilkan oleh aplikasi berbagai metoda *perturbation* pada database bank dapat dilihat pada Tabel 3.

Tabel 3. Karakteristik statistik dari berbagai metoda *perturbation*.

Metoda	Deskripsi	Karakteristik			
		\sum_{ee}	\sum_{XY}	\sum_{YS}	\sum_{YY}
<i>Additive</i>					
SADP	$Y = X + e$	dD	\sum_{XX}	\sum_{XS}	$\sum_{XX} + dD$
CADP	$Y = X + e$	$d \sum_{XX}$	\sum_{XX}	\sum_{XS}	$(1 + d) \sum_{XX}$
BCADP	$Y = [(1/d_1)(X + e)] + [(d_2/d_1)\mu_X]$	$d \sum_{XX}$	$(1/d_1) \sum_{XX}$	$(1/d_1) \sum_{XS}$	\sum_{XX}
<i>Multiplicative</i>					
MDP	$Y = Xe$	dD	\sum_{XX}	\sum_{XS}	$\sum_{XX} + d\mu^2_X D + dD^2$

3.2.1. Metoda Simple Additive Data Perturbation (SADP)

Pada bentuk metoda yang paling sederhana ini, SADP dilakukan dengan melakukan *perturbation* pada atribut konfidensial (X) dengan menambahkan sebuah *noise* (e) untuk menghasilkan nilai atribut yang telah diberi *perturbation* (Y). Jika metoda SADP digunakan untuk database *multi-attribute*, maka setiap atribut dalam database akan diberikan *perturbation* secara independen. SADP dapat dideskripsikan sebagai berikut :

$$Y = X + e \dots\dots\dots (4)$$

dimana e memiliki distribusi normal *multivariate* dengan *mean vector* 0 dan *covariance* matriks dD . Parameter d dan distribusi e dipilih berdasarkan persyaratan yang telah disusun oleh DBA (lihat bagian 3.1). Untuk memenuhi persyaratan (v), ditentukan $d = 1,00$. Persyaratan lainnya tidak akan dipergunakan pada metoda SADP. Hasil pengukuran deskripsi dan keamanan yang dihasilkan dari penggunaan metoda SADP pada database yang telah diberi *perturbation* dapat terlihat pada Tabel 4.

Tabel 4. Pengukuran deskriptif database bank setelah dilakukan metoda *perturbation* SADP.

Summary Statistik		Atribut	Matriks Covariance (Korelasi ditunjukkan dalam kurung)				
Mean (Rp. 000)	Std. Dev. (Rp. 000)		Home Equity	Saham/ Obligasi	Liabilities	Simpanan	Kredit
100,00	28,28	Home Equity	800,00 (1,00)				
50,00	14,14	Saham/ Obligasi	140,00 (0,35)	200,00 (1,00)			
80,00	28,28	Liabilities	320,00 (0,40)	150,00 (0,38)	800,00 (1,00)		
20,00	5,00	Simpanan	50,00 (0,36)	20,00 (0,28)	25,00 (0,18)	25,00 (1,00)	
50,00	10,00	Kredit	60,00 (0,21)	20,00 (0,14)	30,00 (0,11)	30,00 (0,60)	100,00 (1,00)
Keamanan untuk atribut tunggal (S_1) = 1,00			Keamanan untuk kombinasi <i>linear</i> (S_2) = 0,26				

Membandingkan *summary* dari metoda SADP yang termuat pada Tabel 4 dengan *summary* pengukuran data sesungguhnya yang terdapat pada Tabel 2, terlihat bahwa metoda SADP menyebabkan terjadinya *bias* tipe A (*variance* dari atribut yang diberi *perturbation* berbeda dari atribut *originalnya*). Juga terjadi *bias* tipe B, dimana korelasi antar atribut konfidensial sebelum dan sesudah pemberian *perturbation* berbeda. Terjadi juga *bias* tipe C, karena korelasi antara atribut-atribut konfidensial dan atribut-atribut non-konfidensial

yang dihitung sebelum dan sesudah pemberian *perturbation* berbeda. Walaupun demikian *bias* tipe D tidak terjadi, karena baik database *original* maupun *noise* memiliki distribusi normal *multivariate*.

Ditinjau dari perspektif keamanan, terlihat bahwa metoda SADP dapat menyediakan keamanan untuk atribut tunggal secara memadai. Walaupun demikian, keamanan terhadap kombinasi *linear* hanya mencapai 0,26. Ini jauh lebih rendah daripada tingkat keamanan yang diharapkan yaitu 0,50. Metoda SADP perlu diperbaiki untuk meningkatkan keamanan terhadap kombinasi *linear*.

3.2.2. Metoda *Correlated-Noise Additive Data Perturbation (CADP)*

Tidak seperti metoda SADP, metoda CADP menggunakan *correlated-noise* untuk *perturbation*. Metoda CADP dapat ditulis sebagai berikut :

$$Y = X + e \dots\dots\dots (5)$$

dimana e memiliki distribusi normal *multivariate* dengan *mean vector* 0 dan matriks *covariance* $= d \sum_{xx}$. Sebagaimana dalam metoda SADP, persyaratan (iv) menuntut bahwa distribusi e harus normal *multivariate* dan persyaratan (v) menuntut bahwa $d = 1,00$. Hasil dari penggunaan metoda CADP pada database bank termuat pada Tabel 5.

Membandingkan *summary* dari metoda CADP yang termuat pada Tabel 5 dengan *summary* pengukuran data sesungguhnya yang terdapat pada Tabel 2, terlihat bahwa metoda CADP menghasilkan *bias* tipe A dan *bias* tipe C. Tidak terjadi *bias* tipe B dan *bias* tipe D. Metoda CADP menyediakan keamanan yang memadai untuk atribut tunggal. Analisis korelasi *canonical* menunjukkan bahwa keamanan terhadap kombinasi *linear* adalah sebesar 0,39. Ini lebih rendah daripada persyaratan (vi) yang ditetapkan oleh DBA, yaitu 0,50. Hal ini berarti masih terdapat kemungkinan terjadi *partial disclosure* pada metoda CADP jika terdapat atribut non-konfidensial pada database.

Tabel 5. *Pengukuran deskriptif database bank*

setelah dilakukan metoda perturbation CADP.

Summary Statistik			Matriks Covariance (Korelasi ditunjukkan dalam kurung)				
Mean (Rp. 000)	Std. Dev. (Rp. 000)	Atribut	Home Equity	Saham/ Obligasi	Liabilities	Simpanan	Kredit
100,00	28,28	Home Equity	800,00 (1,00)				
50,00	14,14	Saham/ Obligasi	280,00 (0,70)	200,00 (1,00)			
80,00	28,28	Liabilities	640,00 (0,80)	300,00 (0,75)	800,00 (1,00)		
20,00	5,00	Simpanan	50,00 (0,36)	20,00 (0,28)	25,00 (0,18)	25,00 (1,00)	
50,00	10,00	Kredit	60,00 (0,21)	20,00 (0,14)	30,00 (0,11)	30,00 (0,60)	100,00 (1,00)
Keamanan untuk atribut tunggal (S_1) = 1,00				Keamanan untuk kombinasi linear (S_2) = 0,39			

3.2.3. Metoda Bias-Corrected Correlated-Noise Additive Data Perturbation (BCADP)

Modifikasi dari metoda CADP dilakukan untuk mengeliminasi bias tipe A. Metoda BCADP memodifikasi nilai *perturbation* yang dihasilkan dari metoda CADP dengan menggunakan transformasi linear. Secara matematis metoda BCADP dapat dideskripsikan sebagai berikut :

$$Y = (1/d_1)(X + e) + (d_2/d_1)\mu_x \dots\dots\dots (6)$$

dimana $d_1 = (1 + d)^{0,5}$, $d_2 = (d_1 - 1)$, μ_x adalah *mean vector* dari X , dan d adalah level *perturbation* yang diharapkan. Sebagaimana kasus pada metoda CADP, berdasar persyaratan (iv) dan (v), e harus memiliki distribusi normal *multivariate* dengan *mean vector* 0 dan matriks *covariance* $d \sum_{xx}$. Hasil dari aplikasi metoda BCADP ini dapat dilihat pada Tabel 6.

Membandingkan Tabel 2 dan Tabel 6, secara nyata terlihat bahwa bias tipe A, B atau D tidak terjadi pada metoda *perturbation* ini. Namun bias tipe C masih tetap terjadi. Tingkat keamanan yang diberikan oleh metoda BCADP

untuk atribut tunggal hanya sebesar 0,58. Ini lebih rendah dari persyaratan yang ditetapkan, yaitu 1,00. Ini berarti bahwa masih terdapat kemungkinan terjadi *partial disclosure* pada atribut individual. Keamanan yang diberikan terhadap kombinasi *linear* adalah sebesar 0,39 juga lebih rendah dari level yang diinginkan yaitu sebesar 0,50. Ini berarti bahwa *partial disclosure* juga masih mungkin terjadi pada kombinasi *partial*. Jadi metoda BCADP gagal untuk memenuhi kedua persyaratan keamanan yang diharapkan oleh DBA.

Tabel 6. Pengukuran deskriptif database bank setelah dilakukan metoda perturbation BCADP.

Summary Statistik			Matriks Covariance (Korelasi ditunjukkan dalam kurung)				
Mean (Rp. 000)	Std. Dev. (Rp. 000)	Atribut	Home Equity	Saham/ Obligasi	Liabilities	Simpanan	Kredit
100,00	20,00	Home Equity	400,00 (1,00)				
50,00	10,00	Saham/ Obligasi	140,00 (0,70)	100,00 (1,00)			
80,00	20,00	Liabilities	320,00 (0,80)	150,00 (0,75)	400,00 (1,00)		
20,00	5,00	Simpanan	35,35 (0,36)	14,14 (0,28)	17,68 (0,18)	25,00 (1,00)	
50,00	10,00	Kredit	42,43 (0,21)	14,14 (0,14)	21,21 (0,11)	30,00 (0,60)	100,00 (1,00)
Keamanan untuk atribut tunggal (S_1) = 0,58			Keamanan untuk kombinasi <i>linear</i> (S_2) = 0,39				

3.3. Multiplicative Data Perturbation (MDP)

Metoda *multiplicative data perturbation* menggunakan bentuk yang berbeda dari metoda ADP. Untuk atribut konfidensial tunggal *original X*, atribut dengan *perturbation Y* dapat diperoleh dari :

$$Y = Xe \dots\dots\dots (7)$$

dimana *e* memiliki *mean* 1,0 dan *variance* tertentu. Karena *mean e* = 1,0, maka tidak akan terjadi *bias* pada estimasi *mean*. Metoda ADP dikembangkan dari metoda SADP ke metoda yang lebih baik, yaitu metoda CADP dan metoda BCADP yang lebih cocok untuk kasus *multivariate*. Namun demikian, tidak satupun pengembangan ini ditujukan untuk metoda MDP. Karenanya, ketika metoda MDP digunakan untuk memberikan *perturbation* secara *multiple* pada attribute-atribut konfidensial, maka *perturbation* harus diberikan pada setiap atribut secara *independen* (sendiri-sendiri) terpisah dari atribut lainnya. Sebagaimana metoda ADP, persyaratan (*v*) dapat digunakan untuk menentukan tingkat dari *d*. Ini merupakan satu-satunya persyaratan yang digunakan secara langsung pada aplikasi metoda MDP. Hasil dari penggunaan metoda MDP pada database nasabah bank diperlihatkan pada Tabel 7.

Membandingkan hasil dari Tabel 2 dan Tabel 7, dapat dilihat bahwa pada metoda MDP terjadi *bias* tipe A. Penggunaan metoda MDP juga menyebabkan terjadinya *bias* pada pengukuran korelasi, baik antar atribut konfidensial (*bias* tipe B) dan antara atribut konfidensial dan atribut non-konfidensial (*bias* tipe C). Jika metoda MDP diaplikasikan untuk *perturbation* pada database dengan distribusi normal *multivariate*, seperti pada database bank, *bias* tipe D juga terjadi. Tingkat keamanan yang diberikan oleh metoda MDP untuk atribut tunggal adalah sebesar 1,04. Ini lebih tinggi dari level yang disyaratkan yaitu 1,00. Tingkat keamanan untuk kombinasi *linear* hanya sebesar 0,27. Ini berarti lebih rendah dari persyaratan yang ditetapkan DBA, yaitu sebesar 0,50.

Tabel 7. Pengukuran deskriptif database bank

setelah dilakukan metoda perturbation MDP.

Summary Statistik			Matriks Covariance (Korelasi ditunjukkan dalam kurung)				
Mean (Rp. 000)	Std. Dev. (Rp. 000)	Atribut	Home Equity	Saham/ Obligasi	Liabilities	Simpanan	Kredit
100,00	28,57	Home Equity	816,00 (1,00)				
50,00	14,28	Saham/ Obligasi	140,00 (0,35)	204,00 (1,00)			
80,00	28,57	Liabilities	320,00 (0,39)	150,00 (0,37)	816,00 (1,00)		
20,00	5,00	Simpanan	50,00 (0,35)	20,00 (0,28)	25,00 (0,18)	25,00 (1,00)	
50,00	10,00	Kredit	60,00 (0,21)	20,00 (0,14)	30,00 (0,11)	30,00 (0,60)	100,00 (1,00)
Keamanan untuk atribut tunggal (S_1) = 1,04				Keamanan untuk kombinasi linear (S_2) = 0,27			

BAB IV

METODA GENERAL UNTUK ADDITIVE DATA PERTURBATION (GADP)

4.1. Kebutuhan Akan Metoda *Perturbation* Baru

Dengan adanya teknik-teknik analisis data yang maju disertai dengan tersedianya *warehouse* data, studi tentang relasi antar atribut menghasilkan hal-hal penting dalam pengetahuan dan sistem pembuatan keputusan. Jika database diberikan *perturbation* dengan menggunakan metoda-metoda *perturbation* yang telah dibahas diatas, maka berarti keputusan diambil berdasarkan informasi yang *bias* (tidak benar) sehingga dapat menghasilkan konsekuensi yang merugikan. Ada suatu saat dalam organisasi ketika pengambilan keputusan kritis harus didasarkan pada korelasi antar atribut dalam database. Hal ini menuntut nilai korelasi antar atribut harus sama, baik sebelum maupun setelah diberikan *perturbation*. Terjadinya *bias* dalam hal ini akan menyebabkan keputusan yang diambil menjadi tidak tepat. Seperti telah digambarkan pada database bank diatas, penggunaan salah satu metoda *perturbation* yang ada akan menghasilkan estimasi korelasi antar atribut-atribut konfidensial dan non-konfidensial yang *bias*. *Bias* ini tidak dapat dieliminasi.

Metoda *perturbation* yang ada juga membatasi kemampuan DBA untuk melakukan pertimbangan *trade-off* antara *bias* dan keamanan serta untuk menentukan kepastian keamanan yang berhubungan dengan terdapatnya atribut non-konfidensial. Hal-hal diatas menyebabkan timbulnya kebutuhan untuk mengembangkan metoda *perturbation* lainnya yang memiliki kemampuan :

- (1) Mempertahankan relasi yang sama antar semua atribut, baik sebelum maupun sesudah pemberian *perturbation*.

- (2) Mempertimbangkan *trade-off* antara tingkat keamanan dan terjadinya *bias*.
- (3) Mempertimbangkan pengaruh atribut non-konfidensial pada tingkat keamanan.

4.2. Metoda *General Additive Data Perturbation* (GADP)

Pada bagian ini akan dibahas metoda *General Additive Data Perturbation* (GADP) yang secara spesifik didesain untuk mengurangi masalah pada metoda-metoda *perturbation* yang ada. Sesuai dengan notasi pada Daftar Notasi, maka X merepresentasikan atribut p yang berupa data numerik konfidensial dan S merepresentasikan atribut q yang berupa data non-konfidensial dengan distribusi normal *multivariate*. Y merepresentasikan atribut p yang telah diberikan *perturbation*.

Hubungan antara matriks *covariance* X , S dan Y dapat digambarkan sebagai berikut :

$$\Sigma^G = \begin{bmatrix} \Sigma_{XX} & & \\ \Sigma_{XS} & \Sigma_{SS} & \\ \Sigma_{XY} & \Sigma_{SY} & \Sigma_{YY} \end{bmatrix} \dots\dots\dots (8)$$

dimana $U = \{X, S\}$ dengan *mean vector*, $\mu_U^T = [\mu_x \mu_s]$ dan matriks *covariance*

$$\Sigma_{UU} = \begin{bmatrix} \Sigma_{XX} & \\ \Sigma_{XS} & \Sigma_{SS} \end{bmatrix} \dots\dots\dots (9)$$

Maka matriks *covariance* Σ^G dapat dituliskan kembali sebagai :

$$\Sigma^G = \begin{bmatrix} \Sigma_{UU} & \\ \Sigma_{UY} & \Sigma_{YY} \end{bmatrix} \dots\dots\dots(10)$$

dimana $\Sigma_{UY} = \begin{bmatrix} \Sigma_{XY} & \Sigma_{SY} \end{bmatrix}$

4.3. Landasan Metoda GADP

Terdapat $U = c_i$ yang merupakan vektor ke i dari U , misalkan $c_i = [X_{i1}, X_{i2}, \dots, X_{ip}, S_{i1}, S_{i2}, \dots, S_{iq}]$. Untuk contoh database bank (Tabel 1), $c_1 = [120,62, 48,86, 73,72, 22,29, 37,75]$, $c_2 = [120,04, 54,21, 88,65, 18,63, 53,69]$, dan sebagainya. Esensi karakteristik dari metoda GADP adalah pembentukan *conditional random vector* ($Y|U = c_i$) yang membentuk nilai *perturbation* untuk disimpan dalam database. Setiap nilai *perturbation* $Y|U = c_i$ merupakan *conditional value* dari Y dari vektor U . Distribusi $Y|U = c_i$ adalah normal *multivariate* dengan nilai dan *variance* yang diharapkan.

$$E(Y | U = c_i) = \mu_x + \sum_{YU} \left(\sum_{UU} \right)^{-1} (c_i - \mu_U) \dots\dots\dots (11)$$

$$Var(Y | U = c_i) = \sum_{YY} - \sum_{YU} \left(\sum_{UU} \right)^{-1} \sum_{UY} \dots\dots\dots (12)$$

Jika proses *perturbation* telah lengkap, kumpulan dari vektor yang telah diberi *perturbation*, $\{Y|U = c_1, Y|U = c_2, Y|U = c_3, \dots, Y|U = c_n\}$ akan merepresentasikan nilai X yang telah diberi *perturbation*. Dapat dilihat bahwa kumpulan nilai ini memiliki distribusi normal *multivariate* dengan nilai yang diharapkan μ_x dan matriks *covariance* = \sum_{YY} . Selanjutnya, *perturbation* tidak hanya dilakukan berdasar pada atribut konfidensial X saja, tetapi juga pada atribut non-konfidensial S . Hal ini memungkinkan untuk menjaga relasi antara Y dan S .

4.4. Metoda GADP Sebagai Bentuk Umum Dari Metoda *Additive Perturbation*

Salah satu fitur penting dari metoda GADP adalah bahwa metoda ini merepresentasikan hampir semua bentuk umum dari metoda *additive data perturbation*. Hal ini berarti bahwa semua metoda ADP yang ada merupakan kasus spesial dari metoda GADP. Hal ini muncul dari fakta bahwa dalam metoda GADP, DBA dapat dengan leluasa menentukan pilihan matriks *covariance* \sum_{XY} , \sum_{SY} dan \sum_{YY} . Dalam Tabel 3 dapat dilihat bahwa

jika ketiga parameter diatas ditentukan sebagai $\sum_{YY} = \sum_{XX}$,
 $\sum_{XY} = (1/d_1)\sum_{XX}$ dan $\sum_{SY} = (1/d_1)\sum_{SX}$ maka hasil dari *perturbation*
 adalah sama dengan jika menggunakan metoda BCADP. Sama dengan hal
 diatas, metoda GADP dapat juga digunakan untuk mereplikasi dua metoda
 lainnya, yaitu SADP dan CADP dengan sangat baik. Namun demikian, metoda
 GADP tidak hanya terbatas pada struktur *covariance* yang terdapat pada Tabel
 3. Dengan memilih ketiga parameter yang tersedia, metoda GADP dapat
 menyediakan karakteristik *perturbation* yang tidak mungkin disediakan dengan
 metoda ADP. Jadi metoda GADP dapat mencakup semua bentuk umum dari
 berbagai metoda ADP.

4.5. Fleksibilitas Metoda GADP

Sifat alamiah dari metoda GADP adalah peningkatan performa yang
 menyangkut terjadinya *bias* dan fleksibilitas yang lebih besar. Dari uraian pada
 bagian 3, terlihat bahwa tidak satupun metoda ADP yang dapat mengeliminir
 keempat tipe *bias*. Penggunaan metoda GADP pada database yang memiliki
 distribusi normal *multivariate* dapat mengeliminir keempat tipe *bias* dengan
 pemilihan matriks *covariance* \sum_{SY} dan \sum_{YY} secara tepat. Hal ini dapat
 dilakukan dengan pertimbangan bahwa metoda GADP digunakan pada
 database dengan spesifikasi $\sum_{YY} = \sum_{XX}$ dan $\sum_{SY} = \sum_{SX}$.

Hasil *joint* matriks *covariance* X , S dan Y dapat ditulis sebagai berikut :

$$\sum^G = \begin{array}{|c|cc} \hline \sum_{XX} & & \\ \hline \sum_{SX} & \sum_{SS} & \\ \hline \sum_{YX} & \sum_{XS} & \sum_{XX} \\ \hline \end{array} \dots\dots\dots (13)$$

Segitiga atas dari matriks *covariance* diatas menunjukkan atribut
original. Setelah *perturbation*, *user* hanya akan melihat bagian segitiga bawah.

Mengingat bahwa kelas dari atribut adalah berbeda (pada segitiga atas kelasnya adalah X dan S , sedangkan pada segitiga bawah kelasnya adalah S dan Y), maka segitiga bawah merupakan *transpose* dari segitiga atas, oleh karenanya maka matriks-matriks akan identik pada setiap aspeknya. Hal ini memungkinkan semua estimasi yang dihasilkan dari data setelah *perturbation* (segitiga bawah) akan sama dengan estimasi yang ditarik dari data sebelum *perturbation* (segitiga atas). Ini berarti bahwa *bias-bias* tipe A, B dan C dapat dieliminasi. Selanjutnya, jika asumsi distribusi normal *multivariate* terpenuhi, maka *bias* tipe D juga akan tereliminasi.

Pada metoda-metoda *perturbation* ADP, jika DBA menentukan tingkat *perturbation* d , maka berarti DBA menentukan dua hal sekaligus, yaitu *bias* dan keamanan. Reduksi pada *bias* akan selalu disertai dengan reduksi pada tingkat keamanan. Hal ini berbeda dengan metoda GADP, dimana DBA dapat mengeliminasi *bias* sementara keamanan dapat tetap dipertahankan pada tingkat yang memadai.

BAB V

PERFORMA METODA GADP

Pada bagian ini akan dibahas performa terjadinya *bias* dan tingkat keamanan yang dapat disediakan dari metoda GADP yang diaplikasikan pada database bank. Pertama kali akan dilakukan pembahasan tentang implementasi metoda GADP dan selanjutnya akan dilakukan perbandingan performa antara metoda GADP dengan performa metoda *perturbation* lainnya.

5.1. Implementasi Metoda GADP

Untuk mengimplementasikan metoda GADP pada suatu database, DBA dapat menentukan tiga parameter, masing-masing \sum_{YY} , \sum_{SY} dan \sum_{XY} . DBA memiliki fleksibilitas yang lebih besar untuk memilih parameter-parameter yang cocok dengan kebutuhannya.

Untuk memenuhi persyaratan (i) maka setiap atribut harus memiliki *variance* yang sama, baik sebelum maupun sesudah *perturbation*. Dengan kata lain persyaratan (i) menuntut bahwa elemen diagonal \sum_{YY} harus sama dengan elemen diagonal \sum_{XX} . Persyaratan (ii) menentukan bahwa tidak boleh ada *bias* dalam pengukuran korelasi antar atribut-atribut konfidensial. Dengan adanya persyaratan (i) dan (ii) maka secara bersama-sama persyaratan tersebut berarti $\sum_{YY} = \sum_{XX}$. Karena $\sum_{YY} = \sum_{XX}$ maka satu-satunya cara untuk memenuhi persyaratan (iii) adalah dengan menentukan $\sum_{SY} = \sum_{SX}$.

Pemilihan parameter lainnya, yaitu \sum_{XY} ditentukan oleh persyaratan keamanan. Salah satu fitur unik dari metoda GADP adalah bahwa \sum_{XY} dapat dipilih untuk meningkatkan keamanan tanpa mengkompromikan terjadinya *bias*. Hubungan antara X dan S menentukan tingkat keamanan maksimum yang dapat disediakan untuk suatu kasus. Karena S dapat sepenuhnya dilihat, maka *snooper* akan dapat memprediksi nilai X menggunakan S saja. Notasi θ

merepresentasikan korelasi *canonical* antara X dan S . *Snooper* dapat menghitung θ^2 , yaitu perbandingan antara variabilitas kombinasi *linear* X , dengan hanya menggunakan kombinasi *linear* dari S . Ini berarti bahwa θ^2 juga merepresentasikan proporsi minimum variabilitas dari kombinasi *linier* X yang dapat dijelaskan menggunakan kombinasi *linear* dari S dan Y . Karenanya maka proporsi yang tidak dapat dijelaskan, yaitu $(1 - \theta^2)$ merupakan representasi dari tingkat keamanan maksimum yang dapat dicapai.

Pada contoh database bank diatas, $\theta = 0,591$ ($\theta^2 = 0,35$), maka menurut metoda *perturbation*, tingkat keamanan maksimum yang dapat dicapai adalah sebesar 0,65 ($1,00 - 0,35$). Tingkat keamanan maksimum ini dapat dicapai dengan menentukan struktur matriks *covariance* $\sum_{XY} = \theta^2 \sum_{XX}$. Pada contoh database bank, parameter \sum_{XY} telah ditentukan sebagai $0,35 \sum_{XX}$. Maka, metoda GADP menyediakan tingkat keamanan tertinggi yang dapat dicapai untuk kombinasi *linear*.

Implementasi metoda GADP untuk database bank dengan menggunakan parameter diatas secara rinci adalah sebagai berikut.

Langkah 1. Hitung μ_U dan \sum_{UU} (*mean* dan matriks *covariance* dari semua atribut dalam database, baik *konfidensial* maupun *non-konfidensial*)

$$\mu_U^T = [100,0 \quad 50,0 \quad 80,0 \quad 20,0 \quad 50,0]$$

$$\sum_{UU} = \begin{bmatrix} 400,00 & 140,00 & 320,00 & 50,00 & 60,00 \\ 140,00 & 100,00 & 150,00 & 20,00 & 20,00 \\ 320,00 & 150,00 & 400,00 & 25,00 & 30,00 \\ 50,00 & 20,00 & 25,00 & 25,00 & 30,00 \\ 60,00 & 20,00 & 30,00 & 30,00 & 100,00 \end{bmatrix}$$

Langkah 2. Pilih matriks *covariance* \sum_{XY} , \sum_{SY} , \sum_{YY} berdasarkan pertimbangan aspek *bias* dan keamanan. Disini akan dipergunakan persyaratan yang telah ditetapkan oleh DBA pada BAB III sebagai berikut :

$$\sum_{YY} = \begin{bmatrix} 400,0 & 140,0 & 320,0 \\ 140,0 & 100,0 & 150,0 \\ 320,0 & 150,0 & 400,0 \end{bmatrix}$$

$$\sum_{YS} = \begin{bmatrix} 50,0 & 60,0 \\ 20,0 & 20,0 \\ 25,0 & 30,0 \end{bmatrix}$$

$$\sum_{YX} = 0,35 * \sum_{XX} = \begin{bmatrix} 141,5 & 49,6 & 113,3 \\ 49,6 & 35,4 & 53,1 \\ 113,3 & 53,1 & 141,6 \end{bmatrix}$$

Langkah 3. Untuk database nasabah :

$$c_1 = [120,62 \quad 48,68 \quad 73,72 \quad 22,29 \quad 37,75]$$

Hitung $E(Y | U = c_1)$ menggunakan rumus (11) dan $Var(Y | U = c_1)$ menggunakan rumus (12) seperti berikut :

$$E(Y | U = c_1) = \begin{bmatrix} 103,02 \\ 48,50 \\ 75,64 \end{bmatrix}$$

$$Var(Y | U = c_1) = \begin{bmatrix} 285,33 & 96,59 & 247,63 \\ 96,59 & 76,99 & 118,27 \\ 247,63 & 118,27 & 333,75 \end{bmatrix}$$

Langkah 4. Buat observasi random dari sebuah distribusi normal *multivariate* menggunakan *mean* dan *variance*. Observasi ini merepresentasikan observasi yang telah diberi *perturbation*. Dengan menggunakan persyaratan pada BAB III, nilai *perturbation* untuk satu set pertama atribut konfidensial pada contoh database bank [120,62 48,68 73,72] dapat dihitung sebagai berikut :

$$(Y | U = c_1) = [103,30 \quad 59,54 \quad 82,79]$$

Ganti nilai *original* dari atribut konfidensial dari database

$$[120,62 \quad 48,68 \quad 73,72]$$
 dengan nilai

$$(Y | U = c_1) = [103,30 \quad 59,54 \quad 82,79]$$

Langkah 5. Ulangi *Langkah 3* dan *Langkah 4* untuk setiap observasi ($i = 1, \dots, n$).

5.2. Performa *Bias* Dan Keamanan Pada Metoda GADP

Hasil dari implementasi metoda GADP untuk database bank dapat dilihat pada Tabel 8.

5.2.1. *Bias*

Membandingkan pengukuran deskripsi yang ditampilkan pada Tabel 2 dan Tabel 8, dapat terlihat bahwa tidak terjadi perbedaan pada semua aspek pengukuran. Dengan kata lain, jika sebaran database berupa distribusi normal *multivariate*, penggunaan metoda GADP akan menghasilkan estimasi tanpa *bias* pada *summary* atau deskripsi statistik (termasuk *query* yang melibatkan SUMS, PERCENTILES, COUNT dan sebagainya). Tidak akan terdapat *bias* pada pengukuran hubungan antar atribut konfidensial ataupun antara atribut konfidensial dan atribut non-konfidensial. Jadi metoda GADP dapat menghilangkan semua tipe *bias*, baik *bias* tipe A, *bias* tipe B, *bias* tipe C maupun *bias* tipe D.

Tabel 8. Pengukuran deskriptif database bank setelah dilakukan metoda perturbation GADP.

Summary Statistik		Atribut	Matriks Covariance (Korelasi ditunjukkan dalam kurung)				
Mean (Rp. 000)	Std. Dev. (Rp. 000)		Home Equity	Saham/ Obligasi	Liabilities	Simpanan	Kredit
100,00	20,00	Home Equity	400,00 (1,00)				
50,00	10,00	Saham/ Obligasi	140,00 (0,70)	100,00 (1,00)			
80,00	20,00	Liabilities	320,00 (0,80)	150,00 (0,75)	400,00 (1,00)		
20,00	5,00	Simpanan	50,00 (0,50)	20,00 (0,40)	25,00 (0,25)	25,00 (1,00)	
50,00	10,00	Kredit	60,00 (0,30)	20,00 (0,20)	30,00 (0,15)	30,00 (0,60)	100,00 (1,00)
Keamanan untuk atribut tunggal (S_1) = 1,30			Keamanan untuk kombinasi <i>linear</i> (S_2) = 0,65				

5.2.2. Evaluasi Tingkat Keamanan

Untuk database bank, tingkat keamanan terhadap *user* dengan akses terbatas, ditentukan oleh DBA paling sedikit 1,00 sesuai persyaratan (v). Keamanan yang dapat disediakan oleh metoda GADP untuk atribut tunggal adalah 1,30. Ini lebih tinggi dari tingkat yang diharapkan, yaitu 1.0.

DBA juga telah menentukan bahwa tingkat keamanan untuk kombinasi *linear* harus paling sedikit 0,50 sesuai dengan persyaratan (vi). Metoda GADP menyediakan tingkat keamanan terhadap *snooper* dengan akses penuh sebesar 0,65. Ini lebih tinggi dari tingkat yang diharapkan, yaitu sebesar 0,50.

Semua metoda *perturbation* ADP telah gagal memenuhi persyaratan tersebut, sehingga masih memungkinkan terjadinya *partial disclosure*. Hanya metoda GADP yang dapat memberikan jaminan bahwa *partial disclosure* tidak akan terjadi untuk semua tipe *user*, baik yang memiliki akses terbatas maupun akses yang besar.

Kesimpulan utama yang dapat ditarik dari analisis ini adalah bahwa jika akan memberikan *perturbation* pada database yang memiliki distribusi normal *multivariate*, maka metoda GADP seharusnya menjadi metoda *perturbation* yang digunakan, baik atas pertimbangan *bias* yang terjadi maupun tingkat keamanan yang dapat dicapai.

5.3. Performa Metoda GADP Pada Populasi Berdistribusi Log-Normal

Dengan tujuan untuk menggeneralisir performa metoda GADP, maka perlu diadakan penilaian performa pada kondisi jika asumsi distribusi normal *multivariate* tidak terpenuhi. Sebagai catatan, banyak kesimpulan yang ditarik dari metoda GADP tergantung hanya pada matriks *covariance* dan dapat digeneralisir bahkan jika asumsi distribusi normal *multivariate* tidak terpenuhi. Dengan demikian maka dapat ditarik kesimpulan sebagai berikut. Pertama : untuk distribusi berbentuk apapun, penggunaan metoda GADP tidak akan menghasilkan *bias* tipe A, B maupun C. Kedua : untuk distribusi berbentuk apapun, tingkat keamanan yang disediakan metoda GADP, baik untuk atribut tunggal maupun untuk kombinasi *linear* adalah sama. Jadi, hampir semua hasil yang ditunjukkan pada Tabel 4 hingga Tabel 8, tetap benar bahkan jika asumsi distribusi normal *multivariate* tidak terpenuhi. Hanya *bias* tipe D yang akan terpengaruh oleh adanya distribusi *non-normal*.

Untuk data dengan pola distribusi *log-normal*, maka penggunaan *multiplicative perturbation* untuk atribut *log-normally distributed* adalah sama dengan penggunaan *additive perturbation* untuk atribut berdistribusi normal. Karenanya, jika sebaran terdistribusi *log-normal*, maka pendekatannya adalah dengan pertama-tama mentransform data (dengan mengambil algoritmanya), mengimplementasikan metoda GADP untuk data yang telah ditransform, dan kemudian mentransform kembali data ke distribusi *originalnya*.

BAB VI

KESIMPULAN

Tujuan dari berbagai metoda data *perturbation* adalah untuk mengamankan database dari *snooper* ketika menyediakan data secara akurat untuk *user* yang memiliki otorisasi yang sah. Pengembangan metoda *data perturbation* (dari SADP dan MDP ke CADP dan kemudian ke BCADP) dilakukan untuk memenuhi harapan akan peningkatan performa *bias* dan tingkat keamanan yang dapat disediakan. Studi ini memperlihatkan bahwa dengan adanya atribut non-konfidensial, semua metoda *data perturbation* ADP dan MDP memiliki keterbatasan dalam aspek terjadinya *bias* dan tingkat keamanan yang dapat disediakan.

Metoda GADP merupakan metoda data *perturbation* yang terbaru. Metoda ini juga merupakan bentuk umum dari berbagai metoda *additive data perturbation*. Semua metoda ADP merupakan kasus spesial dari metoda GADP. Dalam kondisi distribusi normal *multivariate*, performa metoda GADP lebih baik dibandingkan dengan metoda-metoda *data perturbation* yang lain, termasuk metoda MDP.

Metoda *perturbation* ADP juga membatasi DBA dengan hanya satu parameter, yaitu level *perturbation*. Hal ini menghilangkan kemampuan DBA untuk melakukan *trade-off* antara *bias* dan *keamanan*. Kebalikannya, metoda GADP lebih fleksibel dan memberikan DBA kesempatan untuk menentukan :

- (1) Tingkat *bias* yang dapat diterima untuk *summary* pengukuran.
- (2) Tingkat *bias* yang dapat diterima untuk pengukuran hubungan antar atribut konfidensial.

- (3) Tingkat bias yang dapat diterima untuk pengukuran hubungan antara atribut konfidensial dan atribut non-konfidensial.
- (4) Tingkat keamanan yang dapat disediakan untuk atribut tunggal.
- (5) Tingkat keamanan yang dapat disediakan untuk kombinasi *linear*.

Ditinjau dari sudut pandang aplikasi, spesifikasi-spesifikasi metoda GADP ini seyogyanya tidak hanya perlu untuk diketahui oleh DBA, tetapi juga oleh pihak lain seperti manajer dan pengelola data *warehouse* dalam organisasi, sehingga dapat diterapkan dengan lebih baik.

DAFTAR PUSTAKA

- [1] Cuppen, Menno. *Source Data Perturbation in Statistical Disclosure Control*, <http://neon.vb.cbs.nl/casc/related/sdp.pdf>, 14 Desember 2004, +/- 09.30 WIB.
- [2] Goldstein, Harvey; Myers, Kate. *Freedom of information: towards a code of ethics for performance indicators*, <http://www.mlwin.com/hgpersonal/code-of-ethics-for-performance-indicators.pdf>, 20 Desember 2004, +/- 13.30 WIB.
- [3] George T. Duncan; Sallie A. Keller-McNulty; S. Lynne Stokes. *Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map*, <http://www.niss.org/technicalreports/tr142.pdf>, 15 Desember 2004, +/- 08.30 WIB.
- [4] Josep, Domingo-Ferrer. *Advances in Inference Control in Statistical Databases: An Overview*, <http://neon.vb.cbs.nl/casc/overview.pdf>, 16 Desember 2004, +/- 19.00 WIB.
- [5] Muralidhar, Krishnamurty; Parsa, Rahul; Sarathy, Rathindra. *A General Additive Data Perturbation Method for Database Security*, <http://gatton.uky.edu/faculty/muralidhar/GADP.pdf>, 14 Desember 2004, +/- 09.30 WIB.
- [6] Muralidhar, Krishnamurty; Sarathy, Rathindra. *A theoretical basis for perturbation methods*, <http://gatton.uky.edu/faculty/muralidhar/StatComp.paper.pdf>, 14 Desember 2004, +/- 09.30 WIB.
- [7] Stanley R. M. Oliveira; Osmar R. Za'iane. *Privacy Preserving Clustering By Data Transformation*, <http://www.cs.ualberta.ca/~zaiane/postscript/sbbd03.pdf>, 14 Desember 2004, +/- 09.30 WIB.