การปลอมแปลงฐานข้อมูลอย่างเป็นระบบเพื่อใช้เป็นฮันนี่พอต

นายสิทธิเดช ท้วมพิบูลย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

SYSTEMATIC FALSIFICATION OF DATABASE FOR USING AS HONEYPOT

Mr. Sithidech Tuampiboon

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

หัวข้อวิทยานิพนธ์	การปลอมแปลงฐานข้อมูลอย่างเป็นระบบเพื่อใช้เป็นฮันนี่พอต
โดย	นายสิทธิเดช ท้วมพิบูลย์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.ยรรยง เต็งอำนวย

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.ยรรยง เต็งอำนวย)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.พันธุ์ปิติ เปี่ยมสง่า)

สิทธิเดช ท้วมพิบูลย์ : การปลอมแปลงฐานข้อมูลอย่างเป็นระบบเพื่อใช้เป็นฮันนี่พอต. (SYSTEMATIC FALSIFICATION OF DATABASE FOR USING AS HONEYPOT) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : อ.ดร.ยรรยง เต็งอำนวย, 95 หน้า.

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการปรับเปลี่ยนข้อมูลตัวเลขในฐานข้อมูลอย่างเป็นระบบเพื่อใช้เป็นฮันนี่พอต เพื่อให้ข้อมูลมีความแนบเนียน สามารถลดความตระหนักถึงการรับรู้ว่าข้อมูลที่ผู้โจมตีกำลังมีปฏิสัมพันธ์อยู่นั้นเป็นข้อมูลที่ถูกสร้างขึ้นมาและสามารถรักษาไว้ซึ่งความเป็นส่วนตัวของข้อมูลตัวเลขในฐานข้อมูล ทำให้ข้อมูลต้นฉบับไม่รั่วไหลไปยังผู้โจมตี ผู้วิจัยได้เลือกใช้หลักการของการสุ่มค่าพร้อมกับควบคุมขอบเขตข้อมูลโดยอาศัยข้อมูลต้นฉบับเป็นแนวทาง โดยงานวิจัยนี้ได้นำเสนอวิธีการปรับเปลี่ยนข้อมูล 3 วิธี ได้แก่ การสุ่มค่าข้อมูลโดยควบคุมขอบเขต การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล และการสลับที่ของข้อมูล นอกจากนี้ได้นำเสนอการวิธีปรับเปลี่ยนค่าขอบเขตของข้อมูล ค่าผลรวมของข้อมูลเพื่อปกปิดข้อมูลทางสถิติบางอย่างของข้อมูลต้นฉบับ และได้นำเสนอวิธีการปรับเปลี่ยนรูปแบบการนำเสนอข้อมูล ประกอบด้วยการจัดการค่าว่าง การจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม และการจัดการข้อมูลที่มีรูปแบบตรงกัน โดยผลลัพธ์ของการทดลองงานวิจัยแสดงให้เห็นว่าวิธีการปรับเปลี่ยนข้อมูลที่งานวิจัยนี้นำเสนอ สามารถสร้างข้อมูลที่มีความแนบเนียนและปกปิดข้อมูลต้นฉบับไว้ได้

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2544 .....

5170495021 : MAJOR COMPUTER SCIENCE
KEYWORDS: HONEYPOT / DATA PERTURBATION / PRIVACY-PRESERVATION / RELATIONAL DATABASE

SITHIDECH TUAMPIBOON : SYSTEMATIC FALSIFICATION OF DATABASE FOR USING AS HONEYPOT. ADVISOR : YUNYONG TENG-AMNUAY, Ph.D, 95 pp.

The objective of this research is to propose the systematic falsification of database for using as Honeypot. This will normalize data to avoid the attacker's awareness and preserve the information privacy. This research proposes 3 main techniques. First, methodology for randomization of data including randomization of data by controlled boundary, randomization of data by controlled boundary and distribution of data, and data swapping. Then, methodology for modification data including values which are used as boundary and summary values. Finally, methodology for modifying data presentation including null value management, number of digits after decimal point management, and repetitive pattern of data management. Experimental results indicated that the proposed technique is effective in producing normalized data and in preserving confidentiality of original data.

Department Computer Engineering Student's Signature .....
Field of Study Computer Science Advisor's Signature .....
Academic Year 2011 .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความอนุเคราะห์และความช่วยเหลืออย่างยิ่งจาก อาจารย์ ดร.ยรรยง เต็งอำนวย อาจารย์ที่ปรึกษา ผู้คอยให้คำชี้แนะ ความรู้ แง่คิดที่เป็นประโยชน์ ตลอดจนเป็นผู้ตรวจทานแก้ไขและให้คำเสนอแนะจนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูงที่ช่วยเหลือ ให้โอกาสและความเมตตาแก่ผู้วิจัยเสมอมา

ขอขอบพระคุณ ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และ รองศาสตราจารย์ ดร.พันธุ์ปิติ เปี่ยมสง่า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่คอยให้คำแนะนำในการแก้ไขและปรับปรุงวิทยานิพนธ์ให้มีคุณภาพมากขึ้น รวมไปถึงคณาจารย์และพี่ๆ บุคลากรในภาควิชาที่คอยช่วยเหลือและให้คำแนะนำในการทำวิทยานิพนธ์แก่ผู้วิจัย

และสุดท้ายขอขอบพระคุณบิดา มารดา และครอบครัวที่คอยเป็นกำลังใจให้เสมอมา และขอขอบคุณเพื่อนๆ พี่ๆ และน้องๆ ทุกคน ที่คอยให้กำลังใจและให้คำปรึกษาต่างๆ จนผู้วิจัยสามารถทำวิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วง

สารบัญ

บทที่ หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ความปลอดภัยของข้อมูลเป็นอีกหนึ่งปัจจัยที่หลายองค์กรให้ความสำคัญ เนื่องจากรูปแบบของการโจมตีจากกลุ่มผู้ไม่ประสงค์ดีรวมไปถึงผู้ที่ต้องการแค่ก่อกวนหรือทดสอบความสามารถในการโจมตีได้มีการพัฒนาอย่างไม่หยุดยั้ง ในขณะที่การป้องกันที่ใช้กันทั่วไปเป็นการป้องกันในด้านเดียว ไม่สามารถตอบโต้หรือมีหลักฐานเพียงพอในการเอาผิดกับผู้โจมตีได้ รวมไปถึงการป้องกันส่วนใหญ่ล้วนต้องผ่านกระบวนการที่ก่อให้เกิดความเสียหายก่อนที่จะมีการคิดค้นหาวิธีการในการป้องกันและเผยแพร่ข้อมูลดังกล่าวให้กับสังคมเพื่อนำความรู้นั้นมาใช้ในการป้องกันข้อมูลภายในองค์กรตนเอง จึงเป็นความเสียหายที่เกินความจำเป็น ในช่วง 3-4 ปีที่ผ่านมา Lance Spitzner ได้ออกแบบเทคโนโลยีที่ชื่อว่า ฮันนี่พอต (Honeypot) [1, 2, 3, 4] ซึ่งเป็นเทคโนโลยีที่ช่วยแก้ปัญหาดังกล่าว ถือได้เป็นการปฏิวัติระบบการรักษาความปลอดภัยของข้อมูลอีกครั้งหนึ่ง

ฮันนี่พอตเป็นทรัพยากรคอมพิวเตอร์อย่างหนึ่งที่ออกแบบมาเพื่อให้ถูกโจมตี โดยฮันนี่พอตเดิมทีเป็นทรัพยากรที่ไม่มีคุณค่าในเชิงการใช้งาน ไม่สามารถสร้างผลผลิตให้กับองค์กรและผู้ใช้ทั่วไปไม่สามารถเข้าถึงได้ หากมีการเชื่อมต่อใดเข้าถึงฮันนี่พอต สามารถอนุมานได้ว่าการเชื่อมต่อนั้นเป็นการเชื่อมต่อที่ไม่ประสงค์ดี เมื่อใดก็ตามที่มีการเชื่อมต่อเข้าถึงฮันนี่พอต ฟังก์ชันในการบันทึกข้อมูลก็จะทำงาน และทำการบันทึกข้อมูลทุกรายละเอียดไม่ว่าจะเป็นหมายเลขไอพีแอดเดรส พอร์ท เครื่องมือที่ผู้โจมตีใช้ วิธีการที่ใช้ รวมไปถึงวัตถุประสงค์ แรงจูงใจ เป็นต้น สิ่งเหล่านี้ล้วนเป็นพฤติกรรมที่ผู้โจมตีกระทำกับฮันนี่พอตแทบทั้งสิ้นโดยที่ผู้โจมตีไม่สามารถทราบหรือตระหนักได้ว่าสิ่งที่ตนกำลังมีปฏิสัมพันธ์ด้วยนั้นเป็นระบบปลอมหรือถึงแม้โดนโจมตีเสียหายแล้วก็ไม่เกิดผลเสียหายใดๆ ต่อองค์กร ทำให้องค์กรสามารถได้ข้อมูลเกี่ยวกับการโจมตีนั้นๆ โดยที่ไม่มีความเสียหายเกิดขึ้น ซึ่งแตกต่างจากการได้มาซึ่งข้อมูลการโจมตีแบบเก่า และข้อมูลที่ได้เหล่านี้ก็สามารถนำไปวิเคราะห์และนำไปเพิ่มศักยภาพให้กับระบบรักษาความปลอดภัยของข้อมูลภายในองค์กรให้มีประสิทธิภาพมากยิ่งขึ้น

ปัจจุบันมีการนำฮันนี่พอตมาสร้างเป็นระบบสำเร็จรูปเพื่อความง่ายต่อการนำไปใช้งาน เช่น Honeyd [5] Specter [6] Honeywall CDROM ROO [7, 8] หรือ mwcollect [9] เป็นต้น ทำให้การใช้งานฮันนี่พอตเริ่มแพร่หลายและเป็นที่สนใจในหลายองค์กรอันเนื่องมาจากประสิทธิภาพของตัวเทคโนโลยีและความง่ายในการติดตั้งและใช้งาน แต่ด้วยข้อเสียและข้อจำกัด

บางอย่างของฮันนี่พอต ไม่ว่าจะเป็นความสามารถในการซ่อนตัวที่ยังไม่ได้ประสิทธิภาพเท่าที่ควร หรือจุดบกพร่องต่างๆ ที่เกิดขึ้น จึงมีนักวิจัยหลายท่านได้ทำการพัฒนาฮันนี่พอตเพื่อให้มีประสิทธิภาพมากยิ่งขึ้นทั้งในด้านการซ่อนตัวและการใช้งาน จากฮันนี่พอตดั้งเดิมจึงกลายมาเป็นเครือข่ายของฮันนี่พอตที่ชื่อว่าฮันนี่เน็ต (Honeynet) [10, 11, 12] ที่ช่วยเพิ่มความยืดหยุ่นในการทำงาน โดยองค์กรที่ต้องการใช้ฮันนี่เน็ตสามารถนำระบบใดก็ได้ที่เป็นระบบจริงมาทำเป็นฮันนี่เน็ต ทุกการเชื่อมต่อที่เข้ามายังฮันนี่เน็ตจะถูกบันทึกพฤติกรรมไว้ที่ส่วนฮันนี่วอลล์ (Honeywall) นอกจากฮันนี่เน็ตจะช่วยเพิ่มความยืดหยุ่นในการประยุกต์ใช้ฮันนี่พอตแล้วยังมีประสิทธิภาพในการบันทึกข้อมูลและการซ่อนตัวจากผู้โจมตีที่ดีกว่าเดิม ถึงแม้ฮันนี่เน็ตจะเป็นสถาปัตยกรรมที่มีประสิทธิภาพในการได้มาซึ่งข้อมูลของการโจมตีและผู้โจมตี แต่เนื่องด้วยการที่มีการนำระบบจริงมาเป็นส่วนหนึ่งของฮันนี่เน็ตนั้น ถ้าผู้ติดตั้งขาดความรอบคอบก็อาจก่อให้เกิดความเสียหายกับองค์กรได้ รวมไปถึงความแนบเนียนของตัวฮันนี่พอตเอง (องค์ประกอบย่อยภายในฮันนี่เน็ต) ถ้าขาดความแนบเนียนในการประยุกต์ใช้งานทำให้คุณค่าของฮันนี่เน็ตด้อยลงไปได้

ในช่วงหลายปีที่ผ่านมา ฐานข้อมูลเป็นทรัพยากรอย่างหนึ่งที่มีความเสี่ยงในการถูกโจมตีมากที่สุด เนื่องจากฐานข้อมูลเป็นคลังข้อมูลที่บรรจุข้อมูลที่เป็นความลับขนาดใหญ่ที่สุดในองค์กรเมื่อเทียบกับคลังข้อมูลประเภทอื่น เช่น เครื่องบริการแฟ้ม (File Server) เครื่องบริการอีเมล (E-mail Server) หรือเครื่องบริการเว็บ (Web Server) เป็นต้น นอกจากนี้ผู้ดูแลฐานข้อมูล (Database Administrator) ในหลายองค์กรให้ความสำคัญด้านความปลอดภัยของฐานข้อมูลน้อยมาก โดยผลสำรวจในปี 2008 ระบุว่าผู้ดูแลระบบฐานข้อมูลใช้เวลาน้อยกว่า 5% ในการดูแลด้านความปลอดภัย ทำให้ปริมาณข้อมูลในฐานข้อมูลมีการรั่วไหลเพิ่มขึ้นอย่างต่อเนื่องและมากที่สุดในบรรดาคลังข้อมูลประเภทต่างๆ [13, 14]

จากข้อมูลเบื้องต้นการป้องกันความปลอดภัยให้ฐานข้อมูลจึงเป็นสิ่งจำเป็นสำหรับองค์กรเพื่อความปลอดภัยของความลับภายในองค์กร ด้วยเหตุนี้การนำระบบฐานข้อมูลมาใช้เป็นฮันนี่พอตจึงช่วยสร้างความปลอดภัยให้กับข้อมูลขององค์กรในระดับหนึ่ง โดยที่เมื่อใดก็ตามที่ฐานข้อมูลนี้ถูกโจมตีแล้วพบว่าภายในฐานข้อมูลไม่มีข้อมูลอะไรเลย ทำให้ผู้โจมตีไม่มีแรงจูงใจใดๆ ที่จะมีปฏิสัมพันธ์กับฐานข้อมูลนี้ต่อไป ผลที่ตามมาคือ ข้อมูลที่ได้เกี่ยวกับการโจมตีในครั้งนี้มีอย่างจำกัด นั่นคือ อาจจะได้แค่เพียงข้อมูลไอพีแอดเดรส พอร์ต ของผู้โจมตีเพียงเท่านั้น หรือในกรณีที่ทำการใส่ข้อมูลไว้ภายในฐานข้อมูลที่นำมาใช้เป็นฮันนี่พอต ถ้าหากนำฐานข้อมูลจริงที่องค์กรใช้งานมาบรรจุไว้ ผลที่ตามมาคือ ข้อมูลซึ่งเป็นสิ่งที่สำคัญเป็นอย่างมากขององค์กรรั่วไหลไปยังผู้โจมตี กรณีนี้จึงเป็นสิ่งที่ไม่ควรเกิดเป็นอย่างยิ่ง การใส่ข้อมูลปลอมไว้ในฐานข้อมูล จึงเป็น

วิธีหนึ่งที่ช่วยป้องกันการรั่วไหลของข้อมูลจริงที่เกิดจากการใช้ฮันนี่เน็ตได้ แต่ข้อมูลปลอมที่จะนำมาบรรจุไว้สามารถแบ่งได้เป็น 2 กรณี คือ

1) ข้อมูลปลอมที่ขาดความน่าเชื่อถือ ได้แก่ ข้อมูลที่ไม่มีความสัมพันธ์กับองค์กร เช่น ฐานข้อมูลขององค์กรเกี่ยวกับขายเวชภัณฑ์แต่มีข้อมูลเกี่ยวกับอาวุธยุทโธปกรณ์ หรือข้อมูลที่ไม่อยู่ในขอบเขตของความเป็นไปได้ เช่น พนักงานอายุ 200 ปี เป็นต้น ความไม่สัมพันธ์เหล่านี้ทำให้ผู้โจมตีตระหนักได้ว่ากำลังมีปฏิสัมพันธ์กับของปลอม ทำให้ฮันนี่พอตนี่หมดความน่าสนใจ

2) ข้อมูลปลอมที่มีความน่าเชื่อถือเพียงพอ นั่นคือ ข้อมูลที่มีความสัมพันธ์กับองค์กรและอยู่ในขอบเขตของความเป็นไปได้ แต่ข้อมูลเหล่านั้นไม่ใช่ข้อมูลจริงหรือเป็นข้อมูลจริงแต่ในเชิงความสัมพันธ์ภายในฐานข้อมูลไม่ใช่ความสัมพันธ์จริง เป้าหมายของข้อมูลรูปแบบนี้เพื่อทำให้ผู้โจมตีไม่สามารถตระหนักได้ว่าตนกำลังมีปฏิสัมพันธ์กับข้อมูลปลอม ผู้โจมตีจึงยังดำเนินขั้นตอนการโจมตีฐานข้อมูลต่อไป สิ่งที่ได้คือข้อมูลเกี่ยวกับการโจมตีที่มากขึ้นและมีประสิทธิภาพมากขึ้น โดยที่ไม่ก่อให้เกิดความเสียหายกับข้อมูลขององค์กร

ทางผู้วิจัยจึงนำกรณีตัวอย่างดังกล่าวมาพัฒนาประสิทธิภาพของฮันนี่พอต โดยเสนอการปลอมแปลงข้อมูลในฐานข้อมูลเพื่อนำไปใช้เป็นฮันนี่พอตโดยอาศัยฐานข้อมูลที่องค์กรใช้งานจริงเป็นต้นฉบับ วัตถุประสงค์หลักในการดำเนินการปลอมแปลงข้อมูลมี 2 อย่าง คือ

1) เพื่อลดความตระหนักถึงการรู้ตัวว่าตนเองกำลังมีปฏิสัมพันธ์กับของปลอม ยิ่งข้อมูลมีความแนบเนียนมากเพียงใด ผู้โจมตีก็จะไม่รู้ตัวมากขึ้น ทำให้กิจกรรมและเวลาในการมีปฏิสัมพันธ์กับฐานข้อมูลที่บรรจุด้วยข้อมูลปลอมมีมากขึ้น ส่งผลให้ได้ข้อมูลเกี่ยวกับการโจมตีนั้นมากขึ้นตามไปด้วย

2) การรักษาไว้ซึ่งความลับของข้อมูล เนื่องจากมีการนำข้อมูลจริงมาเป็นต้นฉบับในการปลอมแปลง ต้องมีความมั่นใจได้ว่าข้อมูลต้นฉบับนั้นไม่ถูกทราบโดยผู้โจมตี หรือถ้าทราบจะต้องอยู่ในระดับความปลอดภัยที่ยอมรับได้

งานวิจัยนี้ได้ใช้ฐานข้อมูลของ Microsoft SQL Server 2005 เป็นระบบการจัดการฐานข้อมูล (Database Management System – DBMS) โดยมุ่งเน้นเฉพาะข้อมูลตัวเลขจำนวนเต็ม (Integer) เนื่องจากเป็นประเภทข้อมูลที่สามารถอาศัยวิธีการทางคณิตศาสตร์แก้ไขค่าข้อมูลโดยที่ไม่ทำให้ความน่าเชื่อถือของข้อมูลสูญเสียไปได้

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการปลอมแปลงข้อมูลจำนวนเลขในฐานข้อมูลเพื่อใช้เป็นฮันนี่พอตโดยอาศัยข้อมูลจากฐานข้อมูลที่ใช้จริงเป็นข้อมูลต้นฉบับเพื่อลด

ความตระหนักถึงการรับรู้ว่าเป็นข้อมูลจริงของผู้โจมตีและเพื่อป้องกันความเป็นส่วนตัวของข้อมูลต้นฉบับ

1.3 ขอบเขตของการวิจัย

1) งานวิจัยนี้มุ่งเน้นการรับมือกับผู้โจมตีภายนอกองค์กรเป็นหลัก อาจทำให้ไม่ได้ประสิทธิภาพถ้าหากผู้โจมตีเป็นบุคคลภายในองค์กรที่มีความรู้เกี่ยวกับข้อมูลต้นฉบับ

2) งานวิจัยนี้มุ่งเน้นเฉพาะการปลอมแปลงข้อมูลเชิงตัวเลขจำนวนเต็มเพียงเท่านั้น

3) งานวิจัยนี้ไม่สามารถบอกได้ว่าวิธีการปลอมแปลงข้อมูลที่นำเสนอแต่ละวิธีนั้นดีหรือด้อยกว่าแต่ละวิธีมากน้อยเพียงใด แต่แสดงให้เห็นถึงความแตกต่างของแต่ละวิธี ทั้งนี้การเลือกใช้วิธีใดนั้นขึ้นอยู่กับความต้องการขององค์กร

4) การพัฒนาระบบการทำงานทั้งหมดจะกระทำภายใต้ระบบปฏิบัติการวินโดวส์ (Windows) ใช้ระบบฐานข้อมูลไมโครซอฟต์ซีควลเซิร์ฟเวอร์ 2008 R2 (MS SQL Server 2008 R2) และใช้ภาษาซีชาร์พ (C#) ในกระบวนการพัฒนา

1.4 ขั้นตอนและวิธีดำเนินงานวิจัย

1) ศึกษาทฤษฎีพื้นฐานของฮันนี่พอต ฮันนี่เน็ต โครงสร้างของฐานข้อมูล และหลักการปลอมแปลงข้อมูล

2) ออกแบบขั้นตอนวิธีของการปลอมแปลงฐานข้อมูล

3) พัฒนาเครื่องมือในการวิจัย

4) ทดสอบวิธีการที่นำเสนอ

5) วิเคราะห์ผลการทดลอง

6) สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำวิธีการปลอมแปลงฐานข้อมูลเพื่อนำไปใช้เป็นฮันนี่พอตโดยมีรากฐานอยู่บนฐานข้อมูลเดิมที่ใช้งานจริงที่นำเสนอนี้ไปประยุกต์ใช้ภายในองค์กรเพื่อเพิ่มความปลอดภัยให้กับข้อมูลขององค์กร รวมไปถึงได้ข้อมูลการโจมตีจากผู้โจมตีและสามารถนำข้อมูลดังกล่าวไปวิเคราะห์เพื่อหาวิธีการป้องกันให้กับข้อมูลหรือระบบจริง

1.6 ผลงานตีพิมพ์จากวิทยานิพนธ์

วิทยานิพนธ์นี้ได้รับการตอบรับให้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง “SYSTEMATIC FALSIFICATION OF NUMERICAL DATABASE FOR USING AS HONEYPOT” โดย นายสิทธิเดช ท้วมพิบูลย์ และ ดร.ยรรยง เต็งอำนวย ในงานประชุมวิชาการระดับชาติ “The Seventh National Conference on Computing and Information Technology (NCCIT 2011)” ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพฯ ระหว่างวันที่ 11-12 พฤษภาคม 2554

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีและงานวิจัยที่เกี่ยวข้องกับวิทยานิพนธ์เรื่องนี้ประกอบไปด้วยส่วนของความรู้เบื้องต้นเกี่ยวกับฮันนี่พอต ฮันนี่เน็ต การเขยื้อนข้อมูล และบทวิจารณ์งานวิจัยต่างๆ ที่เกี่ยวข้อง โดยมีรายละเอียดดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ฮันนี่พอต (Honeypot)

2.1.1.1 นิยามของฮันนี่พอต

เดิมการพัฒนาความปลอดภัยบนคอมพิวเตอร์ถูกมุ่งเน้นไปยังการพัฒนากลไกการป้องกันแบบตั้งรับ (Passive Defense) [15] เช่น ไฟร์วอลล์ (Firewall) [16, 17] หรือ ระบบการตรวจจับผู้บุกรุก (IDS หรือ Intrusion Detection System) [18] แต่ฮันนี่พอตแตกต่างออกไป ได้มีคนนิยามความหมายของฮันนี่พอตไว้อย่างหลากหลาย แต่นิยามดั้งเดิมที่นิยามไว้โดย Lance Spitzner ผู้ก่อตั้งโครงการฮันนี่เน็ต ได้ให้นิยามไว้ว่า “ฮันนี่พอตเป็นทรัพยากรระบบสารสนเทศโดยที่คุณค่าของฮันนี่พอตปรากฏในรูปของการไม่มีสิทธิ์เข้าถึงหรือการใช้งานโดยไม่ได้รับอนุญาตของทรัพยากรนั้น” จากนิยามนี้สามารถกำหนดเป็นกฎของฮันนี่พอตได้ 2 ประการ ดังนี้

1) วลี “ทรัพยากรระบบสารสนเทศ” ที่ถูกนำมาใช้ในนิยามนี้เพื่อสื่อให้ทราบว่าสิ่งที่จะนำมาทำเป็นฮันนี่พอตสามารถเป็นทรัพยากรคอมพิวเตอร์ต่างๆ เช่น สถานีงาน (Workstation) เครื่องบริการแฟ้ม (File Server) เครื่องบริการอีเมล (Mail Server) เครื่องพิมพ์ (Printer) อุปกรณ์จัดเส้นทาง (Router) อุปกรณ์เครือข่ายต่างๆ รวมไปถึงระบบเครือข่ายทั้งระบบสามารถนำมาใช้เป็นฮันนี่พอตได้ทั้งหมด

2) ฮันนี่พอตต้องถูกจัดวางไว้ในเส้นทางที่มีโอกาสถูกคุกคามและตัวฮันนี่พอตเองไม่มีมูลค่าการผลิต (Production Value)

2.1.1.2 องค์ประกอบพื้นฐานของฮันนี่พอต

การที่ฮันนี่พอตสามารถทำงานได้อย่างมีประสิทธิภาพได้นั้น ประกอบไปด้วยองค์ประกอบต่างๆ ดังนี้

1) อุปกรณ์เครือข่าย (Network Device Hardware) : เป็นอุปกรณ์ที่ประกอบด้วยไฟร์วอลล์ อุปกรณ์จัดเส้นทาง และสวิตช์ (Switch)

2) เครื่องมือในการเฝ้าสังเกตและการบันทึก (Monitoring/logging Tools) : เป็นส่วนที่ทำการเฝ้าสังเกตและบันทึกข้อมูลต่างๆ ที่ผู้โจมตีมีปฏิสัมพันธ์กับฮันนี่พอต

3) สถานีงานด้านการจัดการ (Management Workstation) : เป็นศูนย์กลางที่ช่วยจัดการข้อมูลต่างๆ ที่ได้จากการเฝ้าสังเกตและบันทึกข้อมูลของฮันนี่พอตแต่ละตัว

4) กลไกการแจ้งเตือน (Alerting Mechanism) : เป็นส่วนที่ช่วยในการแจ้งเตือนให้ผู้ดูแลระบบทราบถึงภัยคุกคามที่กำลังเกิดขึ้นกับฮันนี่พอต

5) ส่วนบันทึกการกดแป้นพิมพ์ (Keystroke Logger) : เป็นส่วนที่ใช้ในการบันทึกคำสั่งต่างๆ ที่ผู้โจมตีพิมพ์เข้ามา

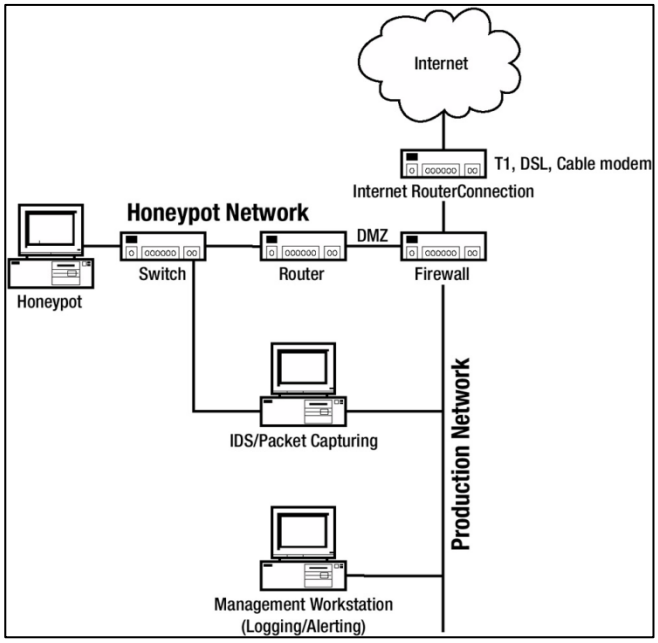
6) ส่วนวิเคราะห์กลุ่มข้อมูล (Packet Analyzer หรือ Sniffer) : เป็นส่วนสำคัญส่วนหนึ่งใช้ในการดักจับกลุ่มข้อมูลที่ผ่านเข้าและออกฮันนี่พอตเพื่อเก็บไว้ตรวจสอบและวิเคราะห์

7) ส่วนการสำรองข้อมูล (Data Backup) : เป็นส่วนที่ใช้ในการสำรองข้อมูลการแก้ไขข้อมูลของผู้โจมตีและสามารถเรียกคืนฮันนี่พอตที่ถูกโจมตีกลับสู่สถานะเดิมที่ไม่มีการโจมตีเกิดขึ้น

8) เครื่องมือจัดการหลักฐาน (Forensic Tools) : เป็นส่วนสำคัญอีกส่วนหนึ่งที่ใช้ในการป้องกันความผิดพลาดในการทำงานหรือการจงใจให้เกิดความเสียหายกับข้อมูลจากผู้โจมตีและยังใช้ค้นหาหลักฐานในกรณีที่เกิดความเสียหายขึ้น ซึ่งมีประโยชน์มากในการติดตามและค้นหาผู้กระทำผิด

9) แหล่งความรู้และข้อมูล (Research Resource) : เป็นอีกส่วนที่ช่วยให้ผู้ดูแลและจัดการฮันนี่พอตสามารถวิเคราะห์ได้ว่าผู้โจมตีได้โจมตีอะไรและเพราะอะไร

นอกจากนี้ ผู้ดูแลระบบเป็นอีกส่วนหนึ่งที่สำคัญซึ่งทำหน้าที่ในการติดตั้งเฝ้าสังเกต และปรับปรุงฮันนี่พอตให้ทันสมัยอยู่เสมอ ตัวอย่างส่วนประกอบของฮันนี่พอตแสดงในรูปที่ 2.1



รูปที่ 2.1 ส่วนประกอบของฮันนี่พอต

2.1.1.3 ชนิดของฮันนี่พอต

ฮันนี่พอตที่ใช้งานทั่วไปสามารถแบ่งประเภทโดยอาศัยเกณฑ์ของระดับปฏิสัมพันธ์ (Interaction Level) ที่เกิดขึ้นระหว่างฮันนี่พอตกับผู้โจมตี สามารถแบ่งได้เป็น 2 ระดับ ดังนี้

1) ฮันนี่พอตระดับปฏิสัมพันธ์ต่ำ (Low-Interaction Honeypot)

ฮันนี่พอตประเภทนี้เป็นประเภทที่มีการติดตั้ง ดูแลรักษา และพัฒนาง่ายที่สุดเนื่องจากมีการออกแบบอย่างง่ายและความสามารถในการทำงานอยู่ในระดับพื้นฐาน รวมไปถึงเซอร์วิซต่างๆ ที่มีไว้เพื่อใช้งานเป็นเพียงเซอร์วิซที่ถูกจำลองขึ้นมาและขอบเขตที่ผู้โจมตีสามารถมีปฏิสัมพันธ์กับฮันนี่พอตได้มีอย่างจำกัด ด้วยลักษณะดังกล่าวทำให้การใช้งานฮันนี่พอตประเภทนี้ก่อให้เกิดความเสี่ยงน้อยที่สุดเมื่อเทียบกับประเภทอื่นๆ เพราะฮันนี่พอตประเภทนี้ไม่มีระบบปฏิบัติการให้ผู้โจมตีสามารถมีปฏิสัมพันธ์ได้ จึงไม่สามารถใช้ฮันนี่พอตโจมตีหรือเฝ้าสังเกตระบบอื่นๆ ที่ใช้งานจริงภายในองค์กร อย่างไรก็ตามด้วยขอบเขตการใช้งานทำให้ข้อมูลเกี่ยวกับการโจมตีที่ได้จากฮันนี่พอตมีอย่างจำกัดด้วย เช่น วันและเวลาในการโจมตี หมายเลขไอพีและพอร์ตต้นทางของการโจมตี หรือหมายเลขไอพีและพอร์ตปลายทางของการโจมตี เป็นต้น ด้วยความง่ายในการใช้งานและข้อจำกัดต่างๆ ฮันนี่พอตประเภทนี้จึงเหมาะสำหรับการศึกษาฮันนี่พอตในขั้นต้น เมื่อมีความเข้าใจเพียงพอแล้วจึงสามารถใช้งานฮันนี่พอตในระดับการมีปฏิสัมพันธ์ที่สูงกว่าต่อไป ตัวอย่างของฮันนี่พอตประเภทนี้ ได้แก่ mwcollect nepenthes และ honeytrap เป็นต้น

2) ฮันนี่พอตระดับปฏิสัมพันธ์สูง (High-Interaction Honeypot)

ฮันนี่พอตประเภทนี้ถูกพัฒนาขึ้นเพื่อให้ผู้โจมตีสามารถมีปฏิสัมพันธ์กับระบบจริง ไม่ใช่ระบบจำลองเหมือนกับฮันนี่พอตระดับปฏิสัมพันธ์ต่ำ ทำให้ผู้โจมตีมีอิสระในการมีปฏิสัมพันธ์กับฮันนี่พอตได้กว้างขึ้น ข้อมูลและความเสี่ยงในการใช้ฮันนี่พอตประเภทนี้จึงมากขึ้นตามไปด้วย กล่าวคือ เนื่องจากผู้โจมตีสามารถมีปฏิสัมพันธ์ได้มากขึ้น ข้อมูลที่ได้เกี่ยวกับการโจมตีนั้นย่อมมากขึ้นตาม เช่น ทราบถึงเครื่องมือที่ใช้ในการโจมตี ข้อความสนทนาระหว่างการโจมตี รวมไปถึงวัตถุประสงค์ของการโจมตี เป็นต้น ในทางตรงกันข้าม ด้วยการใช้ระบบจริงเป็นฮันนี่พอตทำให้ผู้โจมตีสามารถเข้าควบคุมฮันนี่พอตแล้วทำการโจมตีระบบอื่นๆ ที่ใช้งานจริงในองค์กรหรือใช้ฮันนี่พอตในการก่ออาชญากรรมทางคอมพิวเตอร์ด้านอื่นๆ ทำให้ความเสี่ยงในการใช้ฮันนี่พอตประเภทนี้มีมากขึ้นตาม ตัวอย่างของฮันนี่พอตประเภทนี้ที่ได้รับความนิยมคือ ฮันนี่เน็ต ซึ่งจะกล่าวรายละเอียดในหัวข้อถัดไป

2.1.1.4 ข้อดีของฮันนี่พอต

หลังจากที่มีการคิดค้นฮันนี่พอตขึ้นมา นักวิจัยต่างให้ความสนใจในการพัฒนาฮันนี่พอต ทำให้ฮันนี่พอตสามารถประยุกต์ใช้งานได้หลากหลายและมีประสิทธิภาพในการหลอกล่อและเก็บข้อมูลเกี่ยวกับภัยคุกคาม ข้อดีและประโยชน์ของฮันนี่พอต มีดังนี้

1) มีค่าเอฟพีและเอฟเอ็นต่ำ (FP : False-Positive, FN : False-Negative)

ค่าเอฟพีเป็นค่าที่เกิดจากเครื่องมือทางด้านความปลอดภัยบ่งชี้ว่ากิจกรรมที่เป็นภัยคุกคามไม่ใช่ภัยคุกคาม ส่วนค่าเอฟเอ็นเป็นค่าที่เกิดจากเครื่องมือทางด้านความปลอดภัยไม่ได้บ่งชี้ว่าภัยคุกคามเป็นภัยคุกคาม เนื่องจากฮันนี่พอตเป็นทรัพยากรที่ไม่มีมูลค่าการผลิตที่ถูกต้องและไม่มีใครสามารถเข้าถึงฮันนี่พอตได้ยกเว้นผู้ดูแลระบบ ทุกการเชื่อมต่อที่ผ่านเข้าและออกสามารถอนุมานได้ว่าเป็นภัยคุกคาม ดังนั้นข้อมูลที่บันทึกได้จากฮันนี่พอตทุกอย่างต้องถูกนำมาพิจารณาเพราะข้อมูลดังกล่าวต่างเป็นข้อมูลเกี่ยวกับภัยคุกคามทั้งหมด

2) การตรวจจับที่รวดเร็ว (Early Detection)

ด้วยค่าเอฟพีและเอ็นพีที่ต่ำ ทำให้ความเร็วในการตรวจจับภัยคุกคามของฮันนี่พอตเกิดขึ้นอย่างรวดเร็วและแม่นยำ ผู้ดูแลระบบบางคนใช้เทคโนโลยีฮันนี่โทเค็น (Honeytoken) ซึ่งเป็นการนำอ็อบเจกต์ (Object) ที่ไม่มีมูลค่าผลผลิตนำมาวางในฮันนี่พอตหรือระบบงานทั่วไปเพื่อเตือนให้ทราบว่ามีภัยคุกคามเมื่อฮันนี่โทเค็นถูกนำไปใช้งาน ตัวอย่างเช่น ฮันนี่โทเค็นสามารถเป็นบัญชีผู้ใช้ปลอมชื่อว่า Administrator โดยบัญชีผู้ใช้นี้ไม่มีสิทธิ์ในการใช้งานใดๆ โดยควรเปลี่ยนชื่อบัญชีผู้ใช้ Administrator เดิมที่มีสิทธิ์ในการใช้งานทุกอย่างไปเป็นชื่ออื่นที่ไม่มีผล

ต่อการขัดขวางการโจมตีก่อน เมื่อใดที่มีบุคคลพยายามเข้าสู่ระบบด้วยบัญชีผู้ใช้ Administrator ส่วนการแจ้งเตือนจึงเริ่มต้นทำงาน ทำให้ผู้ดูแลระบบสามารถทราบได้ว่าขณะนี้มีผู้ไม่ประสงค์ดีต่อระบบ ทำให้สามารถป้องกันและรับมือได้ทันท่วงที

3) การตรวจจับภัยคุกคามชนิดใหม่ (New Threat Detection)

เนื่องจากทุกการเชื่อมต่อที่ผ่านเข้าออกฮันนี่พอตถือเป็นภัยคุกคาม สำหรับภัยคุกคามที่เคยเกิดขึ้นแล้วสามารถเปรียบเทียบกับข้อมูลของภัยคุกคามที่ได้รับการบันทึกไว้ สำหรับภัยคุกคามใดที่ไม่พบในที่บันทึกไว้ นั่นคือภัยคุกคามชนิดใหม่ที่เกิดขึ้น

4) เรียนรู้เกี่ยวกับผู้โจมตี (Know Your Enemy)

ฮันนี่พอตสามารถบันทึกข้อมูลทุกอย่างที่ผู้โจมตีมีปฏิสัมพันธ์กับฮันนี่พอต เช่น ข้อมูลเครือข่าย เครื่องมือที่ใช้ในการโจมตี บทสนทนา ชุดคำสั่งที่ใช้ รวมไปถึงสามารถเรียนรู้ได้ว่าผู้โจมตีมีจุดประสงค์อะไรที่ทำการโจมตี ข้อมูลเหล่านี้มีประโยชน์ในการพัฒนาอุปกรณ์รักษาความปลอดภัยของระบบภายในองค์กร

5) ข้อมูลที่ได้เป็นหลักฐานทางกฎหมาย (Honeypot As a Forensics Tools)

ด้วยคุณสมบัติการบันทึกข้อมูลเกี่ยวกับภัยคุกคามที่เกิดขึ้นสามารถใช้เป็นหลักฐานในการเอาผิดผู้โจมตีได้ ผู้เอาผิดสามารถนำข้อมูลที่ได้ลำดับเหตุการณ์ภัยคุกคามที่เกิดขึ้นอย่างเป็นขั้นตอนได้

2.1.1.5 ข้อเสียของฮันนี่พอต

ถึงแม้ฮันนี่พอตมีประโยชน์ในการรับมือกับภัยคุกคามที่เกิดขึ้นกับทรัพยากรคอมพิวเตอร์ขององค์กรได้อย่างมีประสิทธิภาพ แต่ระบบต่างๆ ล้วนมีข้อเสีย สำหรับข้อเสียของฮันนี่พอตมีดังนี้

1) การใช้ฮันนี่พอตทำให้สามารถรู้เฉพาะภัยคุกคามที่เกิดขึ้นกับฮันนี่พอตเท่านั้น เมื่อใดที่ผู้โจมตีทำการโจมตีระบบอื่นๆ ที่ไม่ใช่ฮันนี่พอต ฮันนี่พอตไม่สามารถแจ้งเตือนหรือป้องกันได้ และเมื่อผู้โจมตีสามารถรับรู้ได้ว่าระบบที่ตนเองกำลังมีปฏิสัมพันธ์ด้วยเป็นฮันนี่พอต ผู้โจมตีจึงทำการหลีกเลี่ยงฮันนี่พอตแล้วทำการโจมตีระบบที่ใช้งานจริง

2) ฮันนี่พอตที่เกิดข้อผิดพลาดในการตั้งค่าอาจก่อให้เกิดร่องรอยบางอย่างที่ทำให้ผู้โจมตีสามารถรับรู้ได้ว่าระบบที่กำลังมีปฏิสัมพันธ์อยู่เป็นฮันนี่พอต ตัวอย่างเช่น ฮันนี่พอตที่ทำการเลียนแบบตัวบริการเว็บ (Web Server) เมื่อผู้โจมตีทำการเชื่อมต่อมายังฮันนี่พอต ระบบทำการตอบสนองการเชื่อมต่อโดยการส่งข้อความผิดพลาดกลับไปตามปกติ แต่ในบางกรณีที่

ข้อความตอบสนองเกิดความผิดพลาด เช่น สะกดคำผิดจาก length เป็น lenght ทำให้ความผิดพลาดนี้กลายเป็นร่องรอยที่ผู้โจมตีสามารถตระหนักได้ว่าระบบนี้เป็นฮันนี่พอต เป็นต้น

3) เมื่อฮันนี่พอตถูกโจมตีและถูกควบคุมโดยผู้โจมตี ผู้โจมตีสามารถใช้ฮันนี่พอตในการโจมตีระบบอื่นๆ ภายในองค์กร หรือทำการอัพโหลดข้อมูลผิดกฎหมายไปยังโลกภายนอกทำให้องค์กรเกิดความเสี่ยงต่อการละเมิดลิขสิทธิ์ได้

4) ฮันนี่พอตเป็นระบบที่ใช้เวลา ทรัพยากร และความรู้ในการติดตั้ง ดูแลรักษา และพัฒนาเป็นอย่างมาก และเมื่อติดตั้งใช้งานแล้วต้องได้รับการดูแล หมั่นตรวจสอบทั้งก่อนและหลังที่ฮันนี่พอตจะโดนโจมตี หากบกพร่องในการตรวจสอบอาจเป็นผลให้ฮันนี่พอตกลายเป็นสิ่งที่โจมตีระบบคอมพิวเตอร์ในองค์กรเอง

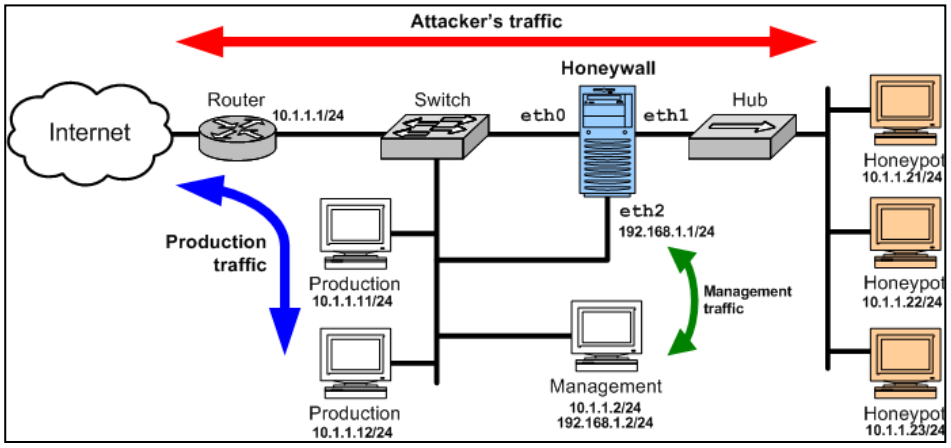
กล่าวโดยสรุป ฮันนี่พอตเป็นทรัพยากรด้านความปลอดภัยประเภทหนึ่งที่ถูกสร้างขึ้นเพื่อให้ถูกโจมตี ตรวจสอบ หรือคุกคาม สามารถเป็นได้ทั้งระบบจริงหรือระบบที่จำลองขึ้นมา หน้าที่หลักคือเป็นสิ่งที่ใช้บันทึกข้อมูลภัยคุกคามที่เกิดขึ้นกับฮันนี่พอตและช่วยป้องกันระบบที่ใช้งานจริงให้มีความปลอดภัยมากยิ่งขึ้น ในปัจจุบันฮันนี่พอตได้รับการพัฒนาอยู่ในรุ่นที่ 2 (GenII) สามารถลดความรุนแรงของภัยคุกคามจากหนักให้เป็นเบาและมีการใช้ตัวบันทึกข้อมูลการพิมพ์เป็นส่วนประกอบหลักในการบันทึกข้อมูล นอกจากนี้เมื่อทำการติดตั้งฮันนี่พอตแล้วต้องหมั่นดูแลและตรวจสอบอยู่เสมอเพื่อความปลอดภัยของระบบคอมพิวเตอร์ภายในองค์กร

2.1.2 **ฮันนี่เน็ต** (Honeynet)

ฮันนี่เน็ตเป็นสถาปัตยกรรมหนึ่ง ที่นำฮันนี่พอตหลายๆ ตัวมาทำงานร่วมกันเป็นเครือข่ายของฮันนี่พอต (**Honey**pot **net**work) ลักษณะสำคัญของฮันนี่เน็ตคือเป็นฮันนี่พอตประเภทระดับปฏิสัมพันธ์สูงที่มีการให้บริการเป็นระบบโปรแกรมประยุกต์หรือเซอร์วิซจริง แตกต่างจากฮันนี่พอตระดับปฏิสัมพันธ์ต่ำที่ให้บริการระบบที่ถูกจำลองขึ้นมา ด้วยความที่ใช้เซอร์วิซจริง จึงสามารถที่จะนำ Microsoft SQL Server 2005 ที่ติดตั้งบนระบบปฏิบัติการ Windows XP หรือ Apache Web Server ที่ติดตั้งบนระบบปฏิบัติการ Ubuntu หรือ Domain Name Server ที่ติดตั้งบนระบบปฏิบัติการ Linux มาทำงานภายใต้เครือข่ายเดียวกันและทำงานร่วมกัน ดังนั้นด้วยความหลากหลายและครอบคลุมที่สามารถนำระบบต่างๆ มาทำงานร่วมกันทำให้ฮันนี่เน็ตสามารถแสดงศักยภาพในการทำงานได้มากกว่าฮันนี่พอตแบบเดิมได้อย่างมาก

ส่วนสำคัญของสถาปัตยกรรมนี้คือฮันนี่วอล (Honeywall) ทำหน้าที่เป็นส่วนควบคุมทุกการเชื่อมต่อทั้งจากภายในและภายนอกฮันนี่เน็ต การที่จะพัฒนาสถาปัตยกรรมนี้ให้มี

ประสิทธิภาพนั้นต้องเริ่มจากการออกแบบและพัฒนาที่ฮันนี่วอลดังตัวอย่างสถาปัตยกรรมฮันนี่เน็ตรุ่นที่ 2 (GenII Honeynet หรือ Generation II Honeynet) ในรูปที่ 2.2



รูปที่ 2.2 ตัวอย่างสถาปัตยกรรมฮันนี่เน็ตรุ่นที่ 2

จากรูปที่ 2.2 สามารถแบ่งองค์ประกอบของฮันนี่เน็ต ออกเป็น 3 ส่วน ได้แก่

1) ส่วนของระบบงานจริงที่ใช้ภายในองค์กร (Production traffic) เป็นส่วนที่ผู้ใช้งานทั่วไปเข้ามาใช้งานเซอร์วิซต่างๆ ในองค์กร โดยทั่วไปแล้วเซอร์วิซเหล่านี้มีระบบรักษาความปลอดภัยในตัวอยู่แล้ว

2) ส่วนของฮันนี่วอล (Honeywall หรือ Management traffic) เป็นส่วนหลักของสถาปัตยกรรมฮันนี่เน็ต ซึ่งจะกล่าวในรายละเอียดต่อไป

3) ส่วนของกลุ่มของฮันนี่พอต เป็นเซอร์วิซจริงต่างๆ ที่ถูกนำมาใช้งานเป็นฮันนี่พอต รูปที่ 2.2 ประกอบด้วย 3 เซอร์วิซ

ทุกการเชื่อมต่อที่เข้าและออกฮันนี่เน็ตต่างต้องผ่านฮันนี่วอล ฮันนี่วอลจึงเป็นส่วนหลักของฮันนี่เน็ตที่ต้องได้รับการออกแบบการทำงานให้มีประสิทธิภาพ โดยส่วนประกอบของฮันนี่วอล ประกอบด้วย 4 ส่วน ได้แก่

1) ส่วนควบคุมข้อมูล (Data Control)

ส่วนนี้เป็นส่วนที่ควบคุมข้อมูลเข้าและออกจากฮันนี่พอต เมื่อผู้โจมตีเข้ามามีปฏิสัมพันธ์กับฮันนี่เน็ตแล้ว ผู้โจมตีมีอิสระในการเชื่อมต่อเข้ามายังฮันนี่เน็ตและสามารถเชื่อมต่อออกไปสู่โลกภายนอกได้ หากไม่มีการควบคุมข้อมูลเข้าและออกดังกล่าวอาจก่อให้เกิดการใช้งานในทางที่ผิด เช่น การใช้ฮันนี่เน็ตโจมตีแบบ DoS (Denial of Service) เป็นต้น ดังนั้นการเลือกจำกัดการเชื่อมต่อออกสู่โลกภายนอกของฮันนี่วอลขึ้นอยู่กับการตัดสินใจของแต่ละองค์กรว่าต้องการได้ข้อมูลในขอบเขตแค่ไหนและพร้อมที่จะแลกกับความเสี่ยงที่เกิดขึ้นมากน้อยแค่ไหนด้วย ในรายงาน

ของโครงการเซาธ์ฟลอริดา (South Florida Project) ได้กล่าวเกี่ยวกับจำนวนการเชื่อมต่อที่เหมาะสมไว้ว่า "*จำนวนการเชื่อมต่อออกสู่ภายนอกจำนวน 5-10 การเชื่อมต่อต่อหนึ่งชั่วโมงนั้นเพียงพอที่จะทำให้ผู้โจมตีพึงพอใจและไม่รู้ตัว ช่วยป้องกันการใช้ฮันนี่เน็ตเป็นเครื่องมือในการค้นหาตรวจตรา หรือโจมตีระบบอื่นได้*"

2) ส่วนบันทึกข้อมูล (Data Capture)

ส่วนนี้ทำหน้าที่ในการเฝ้าสังเกตและบันทึกทุกพฤติกรรมของผู้โจมตีที่เกิดขึ้นภายในฮันนี่เน็ต ข้อมูลที่ได้บอกให้ทราบถึงเครื่องมือ วิธีการ รวมไปถึงแรงจูงใจของผู้โจมตี

3) ส่วนวิเคราะห์ข้อมูล (Data Analysis)

เนื่องจากเป้าหมายหลักของการพัฒนาฮันนี่เน็ตขึ้นมานั้นเพื่อให้ได้มาซึ่งข้อมูล ส่วนนี้จึงเป็นส่วนที่ใช้ในการเปลี่ยนจากข้อมูลดิบที่ได้ให้กลายเป็นสารสนเทศที่มีประโยชน์ ทั้งนี้สารสนเทศที่ได้นั้น แต่ละองค์กรต่างมีความต้องการที่แตกต่างกันไป ในส่วนการวิเคราะห์ข้อมูลนี้จะทำงานได้อย่างมีประสิทธิภาพได้ ผู้ใช้งานต้องตอบคำถามก่อนว่าสารสนเทศที่ต้องการนั้นเป็นอย่างไร มากน้อยเพียงใด หรือมีความละเอียดแค่ไหน

ตารางที่ 2.1 ข้อดีและข้อเสียของฮันนี่เน็ต

ข้อดี	ข้อเสีย
มีความยืดหยุ่น เนื่องจากเป็นการนำระบบหรือเซอร์วิซจริงมาสร้างเป็นเครือข่ายของฮันนี่พอต เพื่อให้ผู้ไม่ประสงค์ดีเข้ามาโจมตี จึงสามารถที่จะนำระบบหรือเซอร์วิซต่างๆ นำมาใช้งาน	มีความซับซ้อนในการใช้งาน การพัฒนาสถาปัตยกรรมฮันนี่เน็ตและใช้ทรัพยากรในการพัฒนาเป็นจำนวนมาก
ข้อมูลที่ได้ครอบคลุมทั้งวิธีการและเครื่องมือของผู้โจมตีทั้งที่เป็นการโจมตีแบบเก่าหรือการโจมตีชนิดใหม่	ก่อให้เกิดความเสี่ยงด้านต่างๆ ดังที่กล่าวไว้ในข้างต้น ถ้าผู้ดูแลระบบขาดความรู้ความเข้าใจอย่างเพียงพอ
สามารถนำไปใช้งานได้กับทุกองค์กรที่ต้องการเพิ่มความปลอดภัยหรือต้องการศึกษาการโจมตีเพื่อนำมาใช้เป็นหลักในการพัฒนาการป้องกันทรัพยากรในองค์กร	เนื่องจากฮันนี่เน็ตเป็นสถาปัตยกรรมที่ใหม่และยังพัฒนาไม่สมบูรณ์ จึงก่อให้เกิดความเสี่ยงใหม่ๆ และปัญหาอื่นๆ ได้

ถึงแม้ดูเหมือนว่าการจะออกแบบและสร้างฮันนี่เน็ตเพื่อมาใช้ในองค์กรเป็นเรื่องที่ยากและใช้เวลา ค่าใช้จ่ายในการออกแบบสูง แต่ทาง Honeynet Project & Research Alliance ได้

ออกแบบสิ่งอำนวยความสะดวกที่รู้จักกันในชื่อของ ฮันนี่วอลซีดีรอม (Honeywall CDROM Roo) เป็นซอฟต์แวร์สำเร็จรูปที่ใช้ในการสร้างฮันนี่เน็ตด้วยความที่ง่ายในการติดตั้ง ปรับแต่ง และใช้งานฮันนี่วอล ทำให้ผู้ใช้ส่วนใหญ่ที่ต้องการสร้างฮันนี่เน็ตไว้ในองค์กรเลือกใช้ฮันนี่วอลซีดีรอม งานวิจัยฉบับนี้จึงเลือกใช้ฮันนี่วอลซีดีรอมเป็นเครื่องมือในการสร้างฮันนี่เน็ตเพื่อทำการวิจัย

2.1.3 **การเขยื้อนข้อมูล** (Data Perturbation) [19, 20, 21, 22, 23]

ฐานข้อมูลเป็นองค์ประกอบหลักอย่างหนึ่งขององค์กร หลายองค์กรจัดเก็บข้อมูลปริมาณมหาศาลไว้ในฐานข้อมูลและข้อมูลส่วนใหญ่เหล่านั้นจัดว่าเป็นความลับขององค์กร ดังนั้นความปลอดภัยของข้อมูลจึงเป็นประเด็นหลักประเด็นหนึ่งที่ต้องพิจารณาในการใช้ระบบฐานข้อมูล

การเขยื้อนข้อมูลเป็นวิธีหนึ่งในการป้องกันข้อมูลลับไม่ให้ถูกเปิดเผย โดยอาศัยการเพิ่มค่ารบกวนอย่างสุ่ม (Random Noise หรือ ε) เข้ากับข้อมูลเชิงจำนวนที่เป็นค่าต้นฉบับ จากนั้นนำผลที่ได้ไปแทนที่ข้อมูลต้นฉบับทำให้ข้อมูลต้นฉบับได้รับการปกปิดไว้

รูปแบบพื้นฐานของการเขยื้อนข้อมูล คือ

$$A' = A + \varepsilon$$

โดยที่ A' คือ ข้อมูลใหม่ที่ได้จากการเขยื้อนข้อมูล

A คือ ข้อมูลต้นแบบ

ε คือ ค่ารบกวนอย่างสุ่ม

ตารางที่ 2.2 ตัวอย่างข้อมูลก่อนและหลังการดำเนินการการเขยื้อนข้อมูล

แถวที่	A		A'		
	ตำแหน่ง	อายุ	ตำแหน่ง	ε	อายุ
1	บัญชี	36	บัญชี	5	41
2	การตลาด	25	การตลาด	2	27
3	การตลาด	27	การตลาด	-9	18
4	บัญชี	33	บัญชี	20	53
5	การตลาด	31	การตลาด	15	46
6	บัญชี	37	บัญชี	27	64

จากตารางที่ 2.2 ข้อมูลที่ผ่านการเขยื้อนข้อมูลแล้วทำให้ได้ข้อมูลใหม่ที่แตกต่างจากข้อมูลต้นฉบับ เป็นการปกปิดข้อมูลต้นฉบับ นอกจากนี้ข้อมูลเชิงสถิติ เช่น ค่าเฉลี่ย ค่าผลรวม ค่าเบี่ยงเบนมาตรฐาน เป็นต้น ต่างเปลี่ยนแปลงไป สามารถรักษาความเป็นส่วนตัวของข้อมูล

2.2 เอกสารและงานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยการนำฮันนี่พอตไปประยุกต์ใช้เป็นฮันนี่ไฟล์ (Honeyfile)

Jim Yuill และคณะ [24] วิจัยเกี่ยวกับการนำฮันนี่พอตไปผนวกเข้ากับแฟ้มข้อมูล โดยทำให้แฟ้มข้อมูลภายในระบบเป็นเสมือนแฟ้มข้อมูลหลอกเพื่อใช้ในการดึงดูดความสนใจให้กับผู้ไม่ประสงค์ดี เช่น สร้างแฟ้มข้อมูลชื่อว่า password.txt ทำให้ผู้ไม่ประสงค์ดีคิดว่า ไฟล์นี้จัดเก็บข้อมูลที่เกี่ยวข้องกับรหัสผ่าน เมื่อใดที่ผู้ไม่ประสงค์ดีทำการเข้าถึงไฟล์นี้ส่วนการแจ้งเตือนทำหน้าที่แจ้งไปยังผู้ดูแลระบบ

2.2.2 งานวิจัยเกี่ยวกับการวัดประสิทธิภาพฮันนี่พอตในการดึงดูดผู้ไม่ประสงค์ดีต่อระบบ

Neil C. Rowe [25] เสนอเกณฑ์ในการวัดประสิทธิภาพฮันนี่พอตในด้านการดึงดูดผู้ไม่ประสงค์ดีต่อระบบ นอกจากนี้ได้มีการสร้างในส่วนของระบบไฟล์ปลอม เพื่อใช้เป็นกรณีศึกษาในการนำเกณฑ์ในการวัดประสิทธิภาพมาใช้งาน

2.2.3 งานวิจัยการประยุกต์ใช้ฮันนี่พอตในลักษณะของฮันนี่พอตฐานข้อมูล (Honeypot Database)

S. K. Gupta และคณะ [26] เสนอสถาปัตยกรรม OCHD (Obliviousness Characteristic of Honeypot Database) ที่รักษาไว้ซึ่งคุณสมบัติของความแนบเนียนของระบบ เป้าหมายหลักของฮันนี่พอตฐานข้อมูล คือ เพื่อเป็นการยืนยันการละเมิดสิทธิ์ของผู้อื่น กล่าวคือ ในการเข้าถึงระบบหนึ่งในองค์กรถึงแม้ระบบดังกล่าวจะมีวิธีการในการยืนยันตัวหรือมีมาตรการรักษาความปลอดภัย แต่ผู้โจมตีสามารถเข้าถึงระบบได้โดยการปลอมแปลงเป็นผู้ที่มีสิทธิ์ในการใช้ระบบ เมื่อเกิดเหตุการณ์เช่นนี้ ฮันนี่พอตฐานข้อมูลทำการสร้างข้อมูลการสอบถามเพื่อใช้ตรวจสอบว่าบุคคลนั้นคือผู้เข้าระบบที่ถูกต้องและมีสิทธิ์ที่แท้จริงหรือไม่

2.2.4 งานวิจัยเกี่ยวกับความปลอดภัยของวิธีการเขยื้อนข้อมูลแบบสุ่ม

M. Krishnamurty และ S.Rathindra ทำการวัดระดับความปลอดภัยของวิธีการเขยื้อนข้อมูล โดยสรุปว่า วิธี Bias Corrected Correlated Noise (BCCN) มีระดับความปลอดภัยที่น้อยกว่าวิธี Correlated Noise (CN) และ Independent Noise (IN)

2.2.5 งานวิจัยเกี่ยวกับผลสำรวจมาตรวัดของขั้นตอนวิธีการทำเหมืองข้อมูลด้านการรักษาความเป็นส่วนตัวของข้อมูล

B. Elisa และคณะ [27] ทำการรวบรวมวิธีและขั้นตอนในการวัดปริมาณความเป็นส่วนตัวของข้อมูล 4 ด้าน ได้แก่

1) ระดับความเป็นส่วนตัว (Privacy Level)

2) ความผิดพลาดในการปกปิด (Hiding Failure)

3) คุณภาพของข้อมูล (Data Quality)

4) ความซับซ้อน (Complexity)

จากความรู้เกี่ยวกับฮันนี่พอต ฮันนี่เน็ต และงานวิจัยต่างๆ ที่เกี่ยวข้องก่อให้เกิดความรู้ที่ช่วยในการพัฒนาวิธีการปรับเปลี่ยนข้อมูลในฐานข้อมูลเพื่อใช้เป็นฮันนี่พอต ช่วยลดความตระหนักของผู้โจมตีว่าตนเองกำลังมีปฏิสัมพันธ์กับข้อมูลปลอมและรักษาความเป็นส่วนตัวของข้อมูลต้นฉบับไว้

บทที่ 3

วิธีดำเนินการวิจัย

ในระบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database) เป็นฐานข้อมูลประเภทหนึ่งที่มีการใช้งานอย่างแพร่หลายในองค์กรเนื่องจากประสิทธิภาพในการจัดการข้อมูลและความง่ายในการใช้ ในฐานข้อมูลประเภทนี้ประกอบด้วยประเภทของข้อมูล (Data Type) หลัก 2 ประเภท คือ ข้อมูลเชิงจำนวน (Numeric Data) และข้อมูลอักขระ (Alphanumeric Strings) สำหรับการปรับเปลี่ยนข้อมูลในฐานข้อมูลเชิงจำนวนสามารถทำได้หลากหลายวิธีโดยนำศาสตร์ในการแปลงข้อมูลอื่นมาใช้ เช่น วิทยาการเข้ารหัสลับ (Cryptography) หลักการเขยื้อนข้อมูล หลักการเพิ่มสิ่งรบกวนให้ข้อมูล (Adding Noise) หลักการแทนที่ข้อมูล (Replacing) หลักการสลับที่ของข้อมูล (Data Swapping) เป็นต้น

ในงานวิจัยนี้มุ่งเน้นเฉพาะการปรับเปลี่ยนข้อมูลเชิงจำนวนเพราะสามารถทำให้แนบเนียนและปกปิดความเป็นส่วนตัวได้ การใช้วิธีทางคณิตศาสตร์เปลี่ยนแปลงข้อมูลเชิงจำนวนนั้นเป็นสิ่งที่ทำได้และตรงกับลักษณะของข้อมูล แต่สำหรับการปรับเปลี่ยนข้อมูลอักขระนั้นซับซ้อนเนื่องจากการเปลี่ยนจากข้อมูลหรือข้อความให้เป็นข้อมูลใหม่ต้องคำนึงถึงองค์ประกอบหลายอย่าง ได้แก่ ความหมายของข้อมูล ความสัมพันธ์ของข้อมูลใหม่และข้อมูลเก่า เป็นต้น ตัวอย่างเช่น ข้อมูล “ห้องปฏิบัติการคอมพิวเตอร์” การหาข้อมูลอื่นมาแทนข้อมูลนั้นได้ต้องพิจารณาก่อนว่าข้อมูลนั้นกล่าวถึงอะไร (ในที่นี้กล่าวถึงสถานที่ที่มีลักษณะเป็นห้อง) ข้อมูลดังกล่าวปรากฏหรือพบเห็นได้ที่ไหน (ในที่นี้พบเห็นได้ในองค์กรที่มีห้องคอมพิวเตอร์จำนวนมากให้ใช้งาน) ด้วยข้อมูลดังกล่าวนี้นำมาพิจารณาแล้วหาสิ่งที่สามารถทดแทนข้อมูลเดิม เช่น ห้องปฏิบัติการเคมี ห้องแม่ข่าย (Server Room) เป็นต้น นอกจากนี้อาจต้องพิจารณาข้อมูลแวดล้อมด้วย

3.1 ลักษณะของข้อมูลหลังการปรับเปลี่ยนข้อมูล

การปรับเปลี่ยนข้อมูลมีประเด็นที่ควรพิจารณา 3 ประการ ดังนี้

3.1.1 ข้อมูลใหม่ต้องมีความแนบเนียนเพียงพอจะลวงให้ผู้โจมตีเข้าใจผิดโดยคิดว่าข้อมูลที่ตนเองกำลังมีปฏิสัมพันธ์นั้นเป็นข้อมูลจริง

3.1.2 ข้อมูลใหม่ต้องสามารถปกปิดข้อมูลต้นฉบับได้ โดยให้ข้อมูลจริงรั่วไหลน้อยที่สุด

3.1.3 ข้อมูลใหม่ต้องมีความแตกต่างในเชิงสถิติกับข้อมูลต้นฉบับ

3.2 วิธีการปรับเปลี่ยนข้อมูลเชิงจำนวน

3.2.1 การสุ่มค่าข้อมูลโดยควบคุมขอบเขต

วิธีการนี้อาศัยหลักของการเขยื้อนข้อมูล เดิมเป็นเพียงการสุ่มค่าเข้ามาหนึ่งค่าแล้วบวกเพิ่มให้กับข้อมูลที่ต้องการปกปิดทำให้ได้ข้อมูลใหม่ที่เปลี่ยนแปลงไปจากเดิม ในบางครั้งข้อมูลใหม่ที่ได้เป็นค่าที่เกินความเป็นจริงหรือขาดความน่าเชื่อถือ เช่น พนักงานอายุ 300 ปี พนักงานมีน้ำหนักตัว 3 กิโลกรัม เป็นต้น งานวิจัยนี้ได้จำกัดข้อมูลใหม่ให้อยู่ในขอบเขตของข้อมูลที่เป็นไปได้โดยอาศัยข้อมูลต้นฉบับที่ใช้งานจริงภายในองค์กรเป็นเกณฑ์ในการพิจารณา สำหรับคำว่า “ขอบเขตของข้อมูลที่เป็นไปได้” หมายถึง ช่วงปิดระหว่างค่าน้อยที่สุดและมากที่สุดที่หาได้จากคอลัมน์ที่กำลังพิจารณาในข้อมูลต้นฉบับ เขียนแทนด้วยสัญลักษณ์ $[min, max] = \{ x \mid min \leq x \leq max\}$ โดยที่ min คือ ค่าน้อยที่สุด max คือค่ามากที่สุด และ x คือค่าข้อมูลใหม่

<u>ตัวอย่างสถานการณ์ในการนำไปใช้</u> สมมติองค์กรหนึ่งต้องการสร้างข้อมูลอายุของพนักงานภายในองค์กรขึ้นมาโดยที่ต้องการให้ข้อมูลที่สร้างขึ้นมานั้นอยู่ภายในช่วงค่าน้อยที่สุดและมากที่สุดของข้อมูลต้นฉบับ คือ 20 ถึง 35 ปี ตามลำดับ โดยไม่สนใจข้อมูลทางสถิติหรือการกระจายของข้อมูล ขอเพียงแค่ข้อมูลใหม่ที่ได้มีความแตกต่างจากข้อมูลต้นฉบับเท่านั้น

การปรับเปลี่ยนข้อมูลดำเนินการทีละ 1 คอลัมน์โดยหาค่าขอบเขตของคอลัมน์ที่กำลังดำเนินการอยู่ เมื่อทำการปรับเปลี่ยนข้อมูลในคอลัมน์ใหม่ต้องทำการหาค่าขอบเขตใหม่ทุกครั้งเสมอ โดยขั้นตอนการสุ่มค่าข้อมูลโดยควบคุมขอบเขต มีดังนี้

กำหนดให้

A คือ ข้อมูลในคอลัมน์

โดยที่ $A = \{A_1, A_2, A_3, \ldots, A_N\}$

A' คือ ข้อมูลในคอลัมน์ที่ผ่านการปรับเปลี่ยนแล้ว

โดยที่ $A' = \{A'_1, A'_2, A'_3, \ldots, A'_N\}$

Min คือ ค่าน้อยที่สุดของชุดข้อมูลในหนึ่งคอลัมน์

โดยที่ $Min = \min\{A_1, A_2, A_3, \ldots, A_N\}$ และ $Min \neq Max$

Max คือ ค่ามากที่สุดของชุดข้อมูลในหนึ่งคอลัมน์

โดยที่ $Max = \max\{A_1, A_2, A_3, \ldots, A_N\}$ และ $Min \neq Max$

ε คือ ค่าที่ทำการสุ่มขึ้นมา

N คือ จำนวนข้อมูลในคอลัมน์

ขั้นตอนที่ 1 จากหลักการของการเขยื้อนข้อมูล

$$A_x' = A_x + \varepsilon_x$$

ขั้นตอนที่ 2 ทำการควบคุมค่าข้อมูลใหม่โดยการใส่ขอบเขตของข้อมูลที่เป็นไปได้ให้สมการในขั้นตอนที่ 1 ทำให้ข้อมูลใหม่ (A_x') อยู่ในช่วง [Min, Max] หรือ

$$A_x' = [Min, Max]$$

เนื่องจาก $A_x' = A_x + \varepsilon_x$ ได้ว่า

$$A_x + \varepsilon_x = [Min, Max]$$

ขั้นตอนที่ 3 ทำการหาขอบเขตของค่า ε_x โดยนำค่า A_x ลบออกจากค่า Min Max และ $A_x + \varepsilon_x$ ได้ว่า

$$\varepsilon_x = [Min - A_x, Max - A_x]$$

ขั้นตอนที่ 4 จากสมการในขั้นตอนที่ 3 ค่า ε_x สามารถเป็นค่า 0 ได้เมื่อบวกเข้ากับข้อมูล A_x แล้วทำให้ข้อมูลใหม่และข้อมูลต้นฉบับเท่ากัน ทำแก้ไขขอบเขตของ ε_x ใหม่เป็น

$$\varepsilon_x = [Min - A_x, Max - A_x] - \{0\}$$

ขั้นตอนที่ 5 เมื่อสุ่มได้ค่า ε_x แล้วนำไปบวกเข้ากับ A_x ได้ว่า

$$A' = \{A_1 + \varepsilon_1, A_2 + \varepsilon_2, A_3 + \varepsilon_3, \ldots, A_N + \varepsilon_N\}$$

การปรับเปลี่ยนข้อมูลแต่ละค่าใน A ต้องทำการหาค่า ε โดยการสุ่มให้อยู่ในขอบเขตของสมการในขั้นตอนที่ 4 จากนั้นนำค่า ε ที่ได้บวกเข้ากับข้อมูลในตำแหน่งนั้นๆ เมื่อดำเนินการครบทุกค่าใน A ทำให้ได้ชุดข้อมูลใหม่ที่ผ่านการปรับเปลี่ยนตามขั้นตอนที่ 5

เพื่อแสดงให้เห็นว่าสมการดังกล่าวช่วยปรับเปลี่ยนข้อมูลจำนวนเลขโดยอยู่ภายในขอบเขตของค่าที่เป็นไปได้ พิจารณาจากตัวอย่างดังนี้

ตัวอย่าง กำหนด A = {5, 6, 7, 8, 9} ได้ขอบเขตของ A = [5, 9]

กรณีที่ค่า $A_1 = 5$ จะได้ช่วงของค่า ε_1 ดังนี้

จาก $\varepsilon_1 = [Min - A_x, Max - A_x] - \{0\}$

แทนค่า จะได้ $\varepsilon_1 = [5 - 5, 9 - 5] - \{0\}$

$\varepsilon_1 = [0, 4] - \{0\}$

$\varepsilon_1 = \{0, 1, 2, 3, 4\} - \{0\}$

$\varepsilon_1 = \{1, 2, 3, 4\}$

แทนค่าในสมการ $A_1' = A_1 + \varepsilon_1$ จะได้ว่า ค่า A_1' สามารถเป็นค่า

- หากสุ่มค่าได้ค่า $\varepsilon_1 = 1$ ทำให้ $A_1' = 5 + 1 = 6$

- หากสุ่มค่าได้ค่า $\varepsilon_1 = 2$ ทำให้ $A_1' = 5 + 2 = 7$
- หากสุ่มค่าได้ค่า $\varepsilon_1 = 3$ ทำให้ $A_1' = 5 + 3 = 8$
- หากสุ่มค่าได้ค่า $\varepsilon_1 = 4$ ทำให้ $A_1' = 5 + 4 = 9$

จะเห็นได้ว่า ค่า A_1' ที่เป็นไปได้ทั้งหมดยังคงอยู่ในขอบเขตของค่าน้อยที่สุดและมากที่สุด หรือ [5, 9]

กรณีค่า $A_5 = 9$ จะได้ช่วงค่า ε_5 ดังนี้

จาก $\varepsilon_5 = [\text{Min} - A_x, \text{Max} - A_x] - \{0\}$

แทนค่า จะได้ $\varepsilon_5 = [5 - 9, 9 - 9] - \{0\}$

$\varepsilon_5 = [-4, 0] - \{0\}$

$\varepsilon_5 = \{-4, -3, -2, -1, 0\} - \{0\}$

$\varepsilon_5 = \{-4, -3, -2, -1\}$

แทนค่าในสมการ $A_5' = A_5 + \varepsilon_5$ จะได้ว่า ค่า A_5' สามารถเป็นค่า

- หากสุ่มค่าได้ค่า $\varepsilon_5 = -4$ ทำให้ $A_1' = 9 + (-4) = 5$
- หากสุ่มค่าได้ค่า $\varepsilon_5 = -3$ ทำให้ $A_1' = 9 + (-3) = 6$
- หากสุ่มค่าได้ค่า $\varepsilon_5 = -2$ ทำให้ $A_1' = 9 + (-2) = 7$
- หากสุ่มค่าได้ค่า $\varepsilon_5 = -1$ ทำให้ $A_1' = 9 + (-1) = 8$

จะเห็นได้ว่า ค่า A_5' ที่เป็นไปได้ทั้งหมดยังคงอยู่ในขอบเขตของค่าน้อยที่สุดและมากที่สุด หรือ [5, 9]

3.2.2 การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล

เมื่อนำข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตพิจารณาการกระจายของข้อมูล พบว่ามีความแตกต่างจากข้อมูลต้นฉบับ ในบางกรณีผู้ทำการปรับเปลี่ยนข้อมูลต้องการให้มีการกระจายของข้อมูลที่เหมือนหรือคล้ายคลึงกับข้อมูลต้นฉบับ การแก้ปัญหาในลักษณะนี้มีงานวิจัยออกมาหลายฉบับ อาทิเช่น Agrawal, R. และ Srikant, R. ที่ได้เสนอวิธีการปรับโครงสร้างของการกระจายของข้อมูลต้นฉบับ เป็นต้น แต่ในงานวิจัยที่ผ่านมายังไม่มีการควบคุมข้อมูลให้อยู่ภายในขอบเขตดังที่ได้กล่าวไว้ในวิธีการสุ่มค่าข้อมูลโดยควบคุมขอบเขต ดังนั้นในงานวิจัยนี้จึงได้นำเสนอวิธีการปรับเปลี่ยนข้อมูลโดยที่ยังคงรักษารูปแบบของการกระจายของข้อมูลและควบคุมขอบเขตของข้อมูลให้อยู่ภายในช่วงค่าที่มีความน่าเชื่อถือโดยอาศัยข้อมูลจากฐานข้อมูลจริงเป็นข้อมูลต้นฉบับ

เนื่องจากในวิธีการสุ่มข้อมูลโดยควบคุมขอบเขตนั้น การนำค่ามากที่สุดและน้อยที่สุดมาเป็นขอบเขตของค่าที่น่าที่เชื่อถือในชุดข้อมูล ทุกครั้งที่ทำการสุ่มค่าข้อมูลใหม่โดยควบคุมให้อยู่ภายในช่วงดังกล่าวจึงสามารถมั่นใจได้ว่าข้อมูลที่ถูกปรับเปลี่ยนมีความน่าเชื่อถือ แต่ทว่าค่ามากที่สุดและน้อยที่สุดที่นำมาใช้เป็นขอบเขตเป็นค่าที่พิจารณาจากทั้งชุดข้อมูลทำให้ช่วงขอบเขตของข้อมูลมีลักษณะที่กว้าง ส่งผลให้การกระจายของข้อมูลที่ผ่านการปรับเปลี่ยนแล้วมีความแตกต่างจากการกระจายของข้อมูลต้นฉบับเป็นอย่างมาก

<u>ตัวอย่างสถานการณ์ในการนำไปใช้</u> องค์กรหนึ่งมีข้อมูลรายรับในช่วงไตรมาสแรกของปี โดยที่ในเดือนมกราคมมีรายรับอยู่ในเกณฑ์สูง เดือนกุมภาพันธ์มีรายรับลดลงเมื่อเทียบกับเดือนมกราคม เดือนมีนาคมมีรายรับเพิ่มขึ้นจากเดือนกุมภาพันธ์แต่น้อยกว่าเดือนมกราคม ทางองค์กรต้องการให้ข้อมูลที่สร้างขึ้นมีลักษณะคล้ายกับรายรับคล้ายกับข้อมูลรายรับในไตรมาสแรกขององค์กรเนื่องจากองค์กรอื่นทราบดีว่าแนวโน้มของรายรับภายในองค์กรเป็นอย่างไรแต่ไม่ทราบตัวเลขรายรับที่แน่นอน ในกรณีนี้ถ้าหากทำการสุ่มค่าโดยไม่ควบคุมการกระจายของข้อมูลทำให้ข้อมูลใหม่ที่ได้มีการกระจายที่แตกต่างไปจากความต้องการขององค์กร ดังนั้นการเลือกใช้การสุ่มค่าโดยควบคุมขอบเขตและการกระจายของข้อมูลจึงสามารถตอบสนองความต้องการขององค์กรได้ในกรณีนี้

ในงานวิจัยนี้ได้เสนอวิธีการควบคุมการกระจายของข้อมูลที่ผ่านการปรับเปลี่ยนแล้วให้มีลักษณะใกล้เคียงกับการกระจายข้อมูลต้นฉบับด้วยการควบคุมขอบเขตของค่าที่สุ่มขึ้นมาให้แคบลงเพื่อให้ข้อมูลใหม่ที่ได้มาที่ค่าใกล้เคียงกับข้อมูลจริง โดยวิธีการมีดังนี้

กำหนดให้

A คือ ข้อมูลในคอลัมน์

โดยที่ $A = \{A_1, A_2, A_3, \ldots, A_N\}$

A' คือ ข้อมูลในคอลัมน์ที่ผ่านการปรับเปลี่ยนแล้ว

โดยที่ $A' = \{A'_1, A'_2, A'_3, \ldots, A'_N\}$

ε คือ ค่าที่ทำการสุ่มขึ้นมา

N คือ จำนวนข้อมูลในคอลัมน์

จากในวิธีการสุ่มค่าข้อมูลโดยควบคุมขอบเขต ค่า ε_x อยู่ภายในขอบเขตของ $[Min - A_x, Max - A_x] - \{0\}$ ในการควบคุมการกระจายของข้อมูลใหม่ วิธีที่นำเสนอในงานวิจัยนี้คือการลดขอบเขตของ ε_x ให้เหลือเพียง $\pm x\%$ ของ A_x โดยที่ค่า x เป็นค่าร้อยละที่ผู้ทำการปรับเปลี่ยนข้อมูลระบุเข้ามาเพื่อจำกัดขอบเขตของ ε_x ให้แคบลงภายในช่วงร้อยละดังกล่าว ทำให้ค่าระหว่างข้อมูลเก่า

และใหม่มีความแตกต่างกันน้อย ส่งผลให้ลักษณะการกระจายของข้อมูลมีความใกล้เคียงมากขึ้น โดยทำการสุ่มค่า ε_x ภายในช่วง ±x% ของ A_x ได้ว่า

จาก $A_x' = A_x + \varepsilon_x$

และ $\varepsilon_x = [-x\%$ ของ $A_x, +x\%$ ของ $A_x] - \{0\}$

ทำการนำ A_x บวกเข้าไปกับทุกค่า

ได้ว่า $A_x' = [A_x - x\%$ ของ $A_x, A_x + x\%$ ของ $A_x] - \{A_x\}$

มีกฎเกณฑ์ในการเลือกค่า $A_x \pm x\%$ ของ A_x ดังนี้

- ถ้า $A_x \pm x\%$ ของ $A_x <$ Min ให้ใช้ค่า Min แทนค่า $A_x \pm x\%$ ของ A_x
- ถ้า $A_x \pm x\%$ ของ $A_x >$ Max ให้ใช้ค่า Max แทนค่า $A_x \pm x\%$ ของ A_x
- สำหรับกรณีอื่นๆ ของ $A_x \pm x\%$ ของ A_x ให้ใช้ค่า $A_x \pm x\%$ ของ A_x

สำหรับกฎเกณฑ์ดังกล่าวมีไว้เพื่อป้องกันไม่ให้ค่า $A_x \pm x\%$ ของ A_x เกินค่าขอบเขตที่เป็นไปได้ โดยที่ถ้าค่าผลรวมมีค่าน้อยกว่าค่าน้อยที่สุดให้ใช้ค่าน้อยที่สุดแทนเพราะถ้าใช้ค่าผลรวมนั้นทำให้ค่าที่ได้มีค่าน้อยกว่าค่าน้อยที่สุด ในทางตรงกันข้าม ถ้าค่าผลรวมมีค่ามากกว่าค่ามากที่สุดให้ใช้ค่ามากที่สุดแทนเพราะถ้าใช้ค่าผลรวมนั้นทำให้ค่าที่ได้มีค่ามากกว่าค่ามากที่สุด

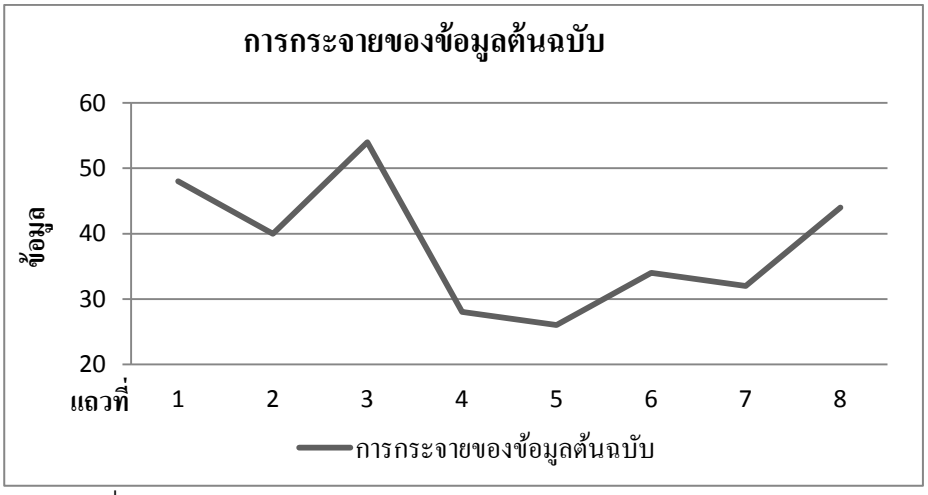
เมื่อคำนวณค่า A_x' ในแต่ละแถวแล้ว นำค่าที่ได้มาแทนที่ค่าในตำแหน่งเดิม ได้เป็น

$$A' = \{A_1', A_2', A_3', \ldots, A_N'\}$$

เพื่อแสดงให้เห็นว่าวิธีการข้างต้นช่วยปรับเปลี่ยนข้อมูลจำนวนเลขโดยอยู่ภายในขอบเขตของค่าที่เป็นไปได้และมีการกระจายของข้อมูลใกล้เคียงกับต้นฉบับ พิจารณาจากตัวอย่างดังนี้

<u>ตัวอย่าง</u> กำหนด A = {48, 40, 54, 28, 26, 34, 32, 44} x = 10%

ข้อมูลนี้ได้ลักษณะการกระจายของข้อมูลดังรูปที่ 3.1 และตัวอย่างข้อมูลในตารางที่ 3.1



รูปที่ 3.1 การกระจายของข้อมูล A = {48, 40, 54, 28, 26, 34, 32, 44}

ตารางที่ 3.1 ตัวอย่างข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล และการสุ่มค่าข้อมูลโดยควบคุมขอบเขต

ข้อมูลที่ (x)	A_x	A_x - x% ของ A_x	A_x + x% ของ A_x	A_x'	**ข้อมูลที่ไม่ควบคุมการกระจาย**
1	48	43	53	45	35
2	40	36	44	41	50
3	54	49	54 59	52	29
4	28	26 25	31	31	38
5	26	26 23	29	28	44
6	34	31	37	31	46
7	32	29	35	35	26
8	44	40	48	47	49

สำหรับข้อมูลที่ 3 ค่าผลรวมของ A_3+ 10% ของ A_3= 59 มีค่ามากกว่าค่ามากที่สุด คือค่า 54 จึงต้องใช้ค่ามากที่สุดแทน

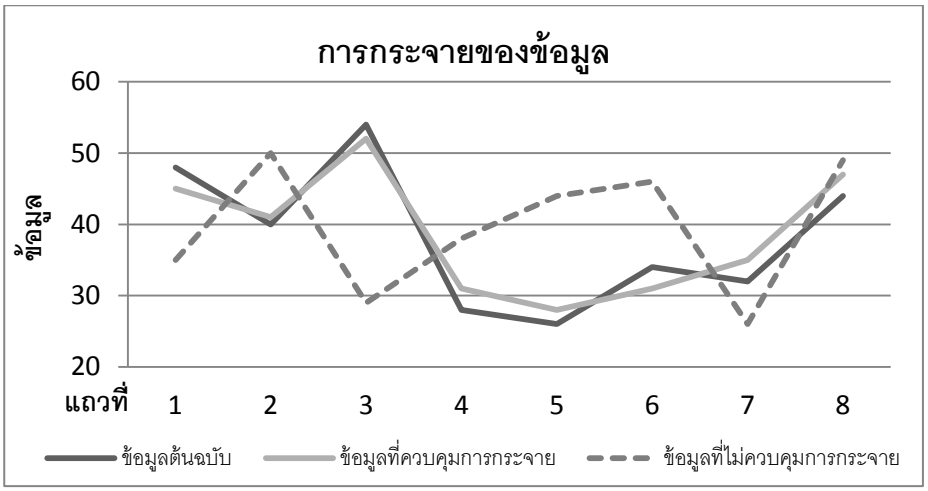
สำหรับข้อมูลที่ 4 ค่าผลรวมของ A_4 - 10% ของ A_4= 25 มีค่าน้อยกว่าค่าน้อยที่สุด คือค่า 26 จึงต้องใช้ค่าน้อยที่สุดแทน

สำหรับข้อมูลที่ 5 ค่าผลรวมของ A_5 - 10% ของ A_5= 23 มีค่าน้อยกว่าค่าน้อยที่สุด คือค่า 26 จึงต้องใช้ค่าน้อยที่สุดแทน

เมื่อแก้ไขและคำนวณเรียบร้อยแล้ว แทนค่าลงใน A′จะได้

A′ = {45, 41, 52, 31, 28, 31, 35, 47}

กราฟที่ได้จากข้อมูล A′เมื่อนำไปสร้างเป็นกราฟได้กราฟดังรูปที่ 3.2



รูปที่ 3.2 การกระจายของข้อมูลต้นฉบับ ข้อมูลหลังการปรับเปลี่ยนโดยวิธีการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล และวิธีการสุ่มค่าข้อมูลโดยควบคุมขอบเขต

จากรูปที่ 3.2 เป็นการเปรียบเทียบการกระจายของข้อมูลระหว่างข้อมูลต้นฉบับ (เส้นทึบสีดำ) กับข้อมูลที่ผ่านการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล (เส้นทึบสีเทา) และข้อมูลที่ผ่านการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตเพียงอย่างเดียว (เส้นประ) พบว่า ลักษณะการกระจายของข้อมูลที่ได้จากกราฟข้อมูลต้นฉบับและข้อมูลที่ผ่านการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูลมีการกระจายของข้อมูลใกล้เคียงกันมากกว่ากราฟระหว่างข้อมูลต้นฉบับและข้อมูลที่ผ่านการสุ่มค่าข้อมูลโดยการควบคุมขอบเขต

3.2.3 **การสลับที่ของข้อมูล** (Data Swapping)

จากสองวิธีที่กล่าวมาข้างต้น สิ่งหนึ่งที่มีการเปลี่ยนแปลงคือค่าข้อมูลทางสถิติ เช่น ค่าผลรวม ค่าเบี่ยงเบนมาตรฐาน การกระจายข้อมูล เป็นต้น ในบางครั้งของการปลอมแปลงข้อมูล ผู้ปลอมแปลงต้องการที่จะรักษาไว้ซึ่งข้อมูลดังกล่าว การแก้ไขข้อมูลจึงทำให้เกิดการเปลี่ยนแปลงของข้อมูลเชิงสถิติ วิธีหนึ่งที่ช่วยรักษาข้อมูลเดิมไว้คือการเปลี่ยนตำแหน่งของข้อมูล กล่าวคือ เดิมข้อมูลอยู่แถวที่ 1 ให้ทำการย้ายตำแหน่งไปอยู่ตำแหน่งที่ 3 เป็นต้น และทุกครั้งหลังจากทำการสลับที่ข้อมูลแล้วต้องทำการตรวจสอบว่าในแต่ละแถวมีข้อมูลต้นฉบับและข้อมูลใหม่เป็นข้อมูลเดียวกัน

หรือไม่ ถ้ามีค่าที่ตรงกันให้ทำการสลับที่ข้อมูลใหม่อีกครั้ง ไปเรื่อยๆ จนกระทั่งไม่มีข้อมูลในแถวใดที่ตรงกัน ผลที่ได้คือลักษณะการกระจายของข้อมูลยังคงเดิม

โดยทั่วไปการสลับตำแหน่งข้อมูลของฐานข้อมูลเชิงสัมพันธ์จะไม่ก่อให้เกิดการเปลี่ยนแปลงข้อมูลในภาพรวม เช่น หากผู้ทำการโจมตีเข้ามาดึงข้อมูลที่ถูกปลอมแปลงด้วยวิธีนี้ไปแล้วทำการจัดเรียงข้อมูลจากมากไปน้อย ผลที่ได้คือข้อมูลหลังการจัดเรียงระหว่างข้อมูลที่ถูกปลอมแปลงและข้อมูลต้นแบบเหมือนกันทุกประการ เป็นต้น แต่เมื่อผู้โจมตีทาการสอบถามข้อมูลจากฐานข้อมูล เช่น ทำการสอบถามว่า พนักงานที่มีรหัสประจาตัว 1150A ได้เงินเดือนกี่บาท จากข้อมูลต้นแบบ ปรากฏข้อมูลคือ รหัส 1150A มีเงินเดือน 50,000 บาท แต่เมื่อทำการสลับกลุ่มข้อมูลแบบสุ่มแล้ว ทาให้ข้อมูลปลอมที่เกิดขึ้นเป็น รหัส 1150A มีเงินเดือน 30,000 บาท ทำให้ผู้โจมตีได้ข้อมูลอื่นที่ไม่ใช่ข้อมูลต้นฉบับ เป็นต้น

<u>ตัวอย่างสถานการณ์ในการนำไปใช้</u> องค์กรหนึ่งต้องการสร้างข้อมูลเงินเดือนใหม่ โดยต้องการรักษาความถี่ของข้อมูลในช่วงต่างๆ เช่น เงินเดือน 5,000 บาทถึง 6,000 บาท มีความถี่ต่ำ เงินเดือน 10,000 บาทถึง 20,000 บาท มีความสูงกว่าเงินเดือนชุดแรก แต่น้อยกว่าช่วงเงินเดือน 30,000 บาทถึง 50,000 บาทซึ่งมีความถี่สูงที่สุด การใช้การสุ่มข้อมูลในวิธีที่ 3.2.1 และ 3.2.2 ไม่สามารถรับประกันได้ว่าข้อมูลใหม่ที่ได้จะมีความถี่ตรงกับข้อมูลต้นฉบับ แต่หากใช้การสลับที่ของข้อมูล ความถี่ของข้อมูลยังคงเป็นเช่นเดิม

3.2.4 การปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุด

วิธีการปรับเปลี่ยนข้อมูล 2 วิธีที่กล่าวไว้ข้างต้นนั้นเป็นการปรับเปลี่ยนข้อมูลโดยที่ข้อมูลทั้งหมดยังคงอยู่ภายในกรอบของขอบเขตที่ใช้ค่ามากที่สุดและน้อยที่สุดของข้อมูลต้นฉบับ แม้ว่าความเป็นส่วนตัวและความน่าเชื่อถือของข้อมูลจะถูกรักษาไว้ แต่สิ่งหนึ่งที่ยังคงไม่ได้รับการปรับเปลี่ยนคือค่าขอบเขตของข้อมูลนั่นเอง ในบางกรณีที่ผู้โจมตีทำการโจรกรรมข้อมูลโดยการสอบถามข้อมูลเกี่ยวกับข้อมูลมากที่สุดหรือน้อยที่สุด

<u>ตัวอย่างสถานการณ์ในการนำไปใช้</u> ในองค์กรหนึ่งเกิดเหตุการณ์ที่ผู้โจมตีทำการสอบถามว่าองค์กรนี้เงินเดือนต่ำสุดที่ให้พนักงานเป็นเท่าไหร่ โดยใช้คำสั่ง SELECT MIN(OrderPrice) AS SmallestOrderPrice FROM Orders เป็นต้น เมื่อผู้โจมตีทำการสอบถามไปยังฐานข้อมูลที่ถูกปรับเปลี่ยนทำให้ได้ข้อมูลที่เป็นค่าน้อยที่สุดเหมือนกับข้อมูลในฐานข้อมูลที่ใช้งานจริงในองค์กร ทำให้ข้อมูลบางอย่างที่เหมือนกับข้อมูลต้นฉบับเกิดการรั่วไหลได้ ดังนั้นค่ามากที่สุดและน้อยที่สุดควรได้รับการปรับเปลี่ยนด้วย

งานวิจัยนี้จึงได้นำเสนอวิธีที่ช่วยหลีกเลี่ยงเหตุการณ์ดังกล่าวด้วยการแก้ไขข้อมูลขอบเขต ในงานวิจัยนี้ได้นำเสนอวิธีการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดด้วยการแทนที่ค่าขอบเขตเดิมด้วยค่าขอบเขตใหม่

วิธีนี้ทำการรับค่ามากที่สุดหรือน้อยที่สุดค่าใหม่จากผู้ทำการปรับเปลี่ยนข้อมูลจากนั้นนำค่าที่รับมาไปแทนที่ค่ามากที่สุดและน้อยที่สุดที่ได้จากข้อมูลต้นฉบับทุกตำแหน่งตามลำดับ โดยที่

กำหนดให้

A คือ ข้อมูลในคอลัมน์

โดยที่ $A = \{A_1, A_2, A_3, \ldots, A_N\}$

A' คือ ข้อมูลในคอลัมน์ที่ผ่านการปรับเปลี่ยนแล้ว

โดยที่ $A' = \{A'_1, A'_2, A'_3, \ldots, A'_N\}$

Min คือ ค่าน้อยที่สุดของชุดข้อมูลในหนึ่งคอลัมน์ และ $Min \neq Max$

โดยที่ $Min = \min\{A_1, A_2, A_3, \ldots, A_N\}$

Max คือ ค่ามากที่สุดของชุดข้อมูลในหนึ่งคอลัมน์ และ $Min \neq Max$

โดยที่ $Max = \max\{A_1, A_2, A_3, \ldots, A_N\}$

Min' คือ ค่าข้อมูลที่รับเข้ามาเพื่อแทนที่ค่า Min โดยที่ $Min' \neq Max'$

Max' คือ ค่าข้อมูลที่รับเข้ามาเพื่อแทนที่ค่า Max โดยที่ $Min' \neq Max'$

N คือ จำนวนข้อมูลในคอลัมน์

ในคอลัมน์หนึ่งมีจำนวน N แถว ซึ่งภายใน A มีค่า Min และ Max เป็นสมาชิก เช่น

$$A = \{A_1, A_2, A_3, \underline{Min}, \underline{Max}, \ldots, A_N\}$$

ทำการแทนที่ค่า *Min* และ *Max* ด้วยค่า Min' และ Max' ตามลำดับ ได้ว่า

$$A' = \{A_1, A_2, A_3, \underline{Min}', \underline{Max}', \ldots, A_N\}$$

เมื่อทำการแทนที่ค่ามากที่สุดและน้อยที่สุดตามที่ผู้ทำการปรับเปลี่ยนข้อมูลได้ระบุเข้ามาแล้ว จึงนำชุดข้อมูลดังกล่าวไปเข้ากระบวนการการสุ่มค่าข้อมูลด้วยการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตหรือการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล ผลที่ได้ของการปรับเปลี่ยนค่าผลรวมคือค่ามากที่สุดและน้อยที่สุดเกิดการเปลี่ยนแปลง ช่วงของข้อมูลที่เป็นไปได้มีการเปลี่ยนแปลงอาจจะมากขึ้นหรือน้อยลงขึ้นอยู่กับค่าที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา

3.2.5 การปรับเปลี่ยนค่าผลรวม

ค่าผลรวมถือเป็นค่าพื้นฐานอย่างหนึ่งของข้อมูลทางสถิติ เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน ค่าสัมประสิทธิ์ของส่วนเบี่ยงเบนเฉลี่ย ค่าสัมประสิทธิ์ของการแปรผัน เป็นต้น นอกจากนี้ค่าสถิติเป็นสิ่งหนึ่งที่ช่วยให้สามารถเข้าใจลักษณะข้อมูลในภาพรวม ดังนั้นการปรับเปลี่ยนค่าผลรวมจึงเป็นวิธีหนึ่งที่ช่วยปกปิดข้อมูลทางสถิติดังกล่าวได้

ตัวอย่างสถานการณ์ในการนำไปใช้ ในองค์กรหนึ่งเกิดเหตุการณ์ที่ผู้โจมตีทำการสอบถามว่าองค์กรนี้มียอดการสั่งซื้อรวมเป็นเท่าไหร่โดยใช้คำสั่ง SELECT SUM(OrderPrice) AS OrderTotal FROM Orders เป็นต้น เมื่อผู้โจมตีทำการสอบถามไปยังฐานข้อมูลที่ถูกปรับเปลี่ยนทำให้ได้ข้อมูลที่เป็นค่าผลรวมที่เหมือนกับข้อมูลในฐานข้อมูลที่ใช้งานจริงในองค์กร ทำให้ข้อมูลบางอย่างที่เหมือนกับข้อมูลต้นฉบับเกิดการรั่วไหลได้ ดังนั้นผลรวมควรได้รับการปรับเปลี่ยนด้วยวิธีการปรับเปลี่ยนค่าผลรวมที่นำเสนอในงานวิจัยประกอบด้วย 2 วิธี ดังนี้

กำหนดให้

A คือ ข้อมูลในคอลัมน์
โดยที่ $A = \{A_1, A_2, A_3, \ldots, A_N\}$

A' คือ ข้อมูลในคอลัมน์ที่ผ่านการปรับเปลี่ยนแล้ว
โดยที่ $A' = \{A'_1, A'_2, A'_3, \ldots, A'_N\}$

sum คือ ค่าผลรวมของข้อมูล A

sum' คือ ค่าผลรวมที่ระบุเข้ามาเพื่อแทนที่ค่า sum โดยที่ $sum \neq sum'$

$\bar{x}$ คือ ค่าเฉลี่ยของข้อมูลต้นฉบับ

$\bar{x}'$ คือ ค่าเฉลี่ยของค่าผลรวมที่ระบุเข้ามา

$\Delta\bar{x}$ คือ ผลต่างระหว่างค่าเฉลี่ยของข้อมูลต้นฉบับและค่าเฉลี่ยของค่าผลรวมที่ระบุ

N คือ จำนวนข้อมูลในคอลัมน์

3.2.5.1 การปรับเปลี่ยนค่าผลรวมกรณีไม่ต้องการเก็บรักษาค่าขอบเขตเดิมไว้ ขั้นตอนประกอบด้วย

ขั้นตอนที่ 1 ทำการหาค่าผลรวมจากชุดข้อมูล A จากสมการ

$$sum = \sum_{i=1}^{N} A_i$$

ขั้นตอนที่ 2 ทำการหาค่าเฉลี่ยของข้อมูลต้นฉบับ จากสมการ

$\bar{x}$ = sum / N

ขั้นตอนที่ 3 ทำการหาค่าเฉลี่ยของค่าผลรวมที่ระบุเข้ามา จากสมการ

$\bar{x}'$ = sum$'$ / N

ขั้นตอนที่ 4 หาผลต่างระหว่างค่าเฉลี่ยของข้อมูลต้นฉบับและค่าเฉลี่ยของค่าผลรวมที่ระบุ จากสมการ

$\Delta\bar{x} = |\bar{x} - \bar{x}'|$

ขั้นตอนที่ 5 นำค่า $\Delta\bar{x}$ บวกเพิ่มหรือลบออกให้กับสมาชิกทุกค่าใน A โดยที่

กรณี sum < sum$'$

A$'$ = {$A_1 + \Delta\bar{x}$, $A_2 + \Delta\bar{x}$, $A_3 + \Delta\bar{x}$, …, $A_N + \Delta\bar{x}$}

กรณี sum > sum$'$

A$'$ = {$A_1 - \Delta\bar{x}$, $A_2 - \Delta\bar{x}$, $A_3 - \Delta\bar{x}$, …, $A_N - \Delta\bar{x}$}

ข้อมูลหลังจากการปรับเปลี่ยนที่ได้แบ่งออกเป็น 2 กรณี กรณีแรกคือกรณีที่ค่าผลรวมที่ระบุเข้ามามากกว่าค่าผลรวมที่ได้จากข้อมูลต้นฉบับ ทำให้ข้อมูลหลังจากการปรับเปลี่ยนทุกค่ามีค่ามากขึ้นตามค่า $\Delta\bar{x}$ ในทางกลับกัน กรณีที่ค่าผลรวมที่ระบุเข้ามาน้อยกว่าค่าผลรวมที่ได้จากข้อมูลต้นฉบับ ทำให้ข้อมูลหลังจากการปรับเปลี่ยนทุกค่ามีค่าลดลงตามค่า $\Delta\bar{x}$

เพื่อแสดงให้เห็นว่าสมการดังกล่าวช่วยปรับเปลี่ยนค่าผลรวมของข้อมูลหลังการปรับเปลี่ยนให้เป็นไปตามที่ผู้ปรับเปลี่ยนต้องการ พิจารณาจากตัวอย่างดังนี้

กำหนด A = {48, 40, 54, 28, 26, 34, 32, 44} และ sum$'$ = 500

ขั้นตอนที่ 1 ทำการหาค่าผลรวมของข้อมูล A

sum = 48 + 40 + 54 + 28 + 26 + 34 + 32 + 44 = 306

ขั้นตอนที่ 2 ทำการหาค่าเฉลี่ยของข้อมูล A

$\bar{x}$ = 306 / 8 = 38.25

ขั้นตอนที่ 3 ทำการหาค่าเฉลี่ยของค่าผลรวมที่ระบุเข้ามา

$\bar{x}'$ = 500 / 8 = 62.5

ขั้นตอนที่ 4 หาผลต่างระหว่างค่าเฉลี่ยของข้อมูลต้นฉบับและค่าเฉลี่ยของค่าผลรวมที่ระบุ

$\Delta\bar{x}$ = |38.25 – 62.5| = 24.25

ขั้นตอนที่ 5 นำค่า $\Delta\bar{x}$ บวกเพิ่มหรือลบออกให้กับสมาชิกทุกค่าใน A

เนื่องจาก sum < sum′

A′ = {48 + 24.25, 40 + 24.25, 54 + 24.25, 28 + 24.25, 26 + 24.25,
34 + 24.25, 32 + 24.25, 44 + 24.25}

A′ = {72.25, 64.25, 78.25, 52.25, 50.25, 58.25, 56.25, 68.25}

ทดสอบหาผลรวมของ A′

sum(A′) = 72.25 + 64.25 + 78.25 + 52.25 + 50.25 + 58.25 +
56.25 + 68.25

= 500

เห็นได้ว่าค่าผลรวมใหม่ที่ได้เป็นไปตามที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา

3.2.5.2 การปรับเปลี่ยนค่าผลรวมกรณีที่ต้องการเก็บรักษาค่าขอบเขตเดิมไว้

จากขั้นตอนของกรณีที่ไม่ต้องการเก็บรักษาค่าขอบเขตพบว่า กรณีที่ค่าผลรวมใหม่มากกว่าผลรวมเก่า ข้อมูลทุกค่าในชุดข้อมูลมีค่าเพิ่มขึ้น ทำให้ไม่สามารถเก็บรักษาค่ามากที่สุดไว้ได้เนื่องจากมีความเป็นไปได้ที่ข้อมูลที่ไม่ใช่ค่ามากที่สุดเมื่อได้รับการเพิ่มค่าเข้าไปแล้วส่งผลให้ค่าดังกล่าวมีค่ามากกว่าค่ามากที่สุดเดิม ในทางตรงกันข้ามค่าน้อยที่สุดเป็นค่าที่สามารถเก็บรักษาไว้ได้เนื่องจากเมื่อค่าอื่นๆ ได้รับการเพิ่มขึ้น ไม่มีทางเป็นไปได้ที่ค่าดังกล่าวน้อยกว่าค่าน้อยที่สุด

ในกรณีที่ค่าผลรวมใหม่น้อยกว่าผลรวมเก่า ข้อมูลทุกค่าในชุดข้อมูลมีค่าลดลง ดังนั้นด้วยเหตุผลในทำนองเดียวกับกรณีแรก สามารถเก็บรักษาค่ามากที่สุดไว้ได้

ขั้นตอนในการปรับเปลี่ยนค่าผลรวมประกอบด้วย

ขั้นตอนที่ 1 ทำการดึงค่ามากที่สุด (กรณี Sum′ < Sum) หรือค่าน้อยที่สุด (กรณี Sum′ > Sum) ออกมาจากชุดข้อมูล เช่น ชุดข้อมูลประกอบด้วย {6, 7, 8, 9, 10} กำหนด Sum′= 60 เนื่องจาก Sum′ > Sum ดังนั้นจึงต้องดึงค่าน้อยที่สุดออกมา ทำให้ได้ชุดข้อมูลใหม่เป็น {7, 8, 9, 10}

ขั้นตอนที่ 2 ทำการหาค่าผลรวมของชุดข้อมูลที่ผ่านการดึงค่าขอบเขตออกมาแล้ว จากตัวอย่างข้างต้น ได้ผลรวมใหม่เป็น Sum{7, 8, 9, 10} = 34

ขั้นตอนที่ 3 ทำการหาผลรวมของข้อมูลที่ถูกดึงออกไปจากชุดข้อมูลเดิม จากตัวอย่างข้างต้น ได้ว่า Sum{6} = 6

ขั้นตอนที่ 4 ทำการหาค่าสัมบูรณ์ของผลต่างระหว่างค่าผลรวมใหม่ที่ระบุเข้ามากับค่าที่ได้จากขั้นตอนที่ 3 จากตัวอย่างข้างต้น ได้ว่า |60 – 6| = 54

ขั้นตอนที่ 5 ทำการหาค่าเฉลี่ยของค่าที่ได้จากขั้นตอนที่ 2 โดยการนำค่าที่ได้จากขั้นตอนที่ 2 หารด้วยจำนวนข้อมูลที่ผ่านการดึงค่าขอบเขตออกมาแล้ว จากตัวอย่างข้างต้น ได้ว่า 34/4 = 8.5

ขั้นตอนที่ 6 ทำการหาค่าเฉลี่ยของค่าที่ได้จากขั้นตอนที่ 4 โดยการนำค่าที่ได้จากขั้นตอนที่ 4 หารด้วยจำนวนข้อมูลที่ผ่านการดึงค่าขอบเขตออกมาแล้ว จากตัวอย่างข้างต้น ได้ว่า 54/4 = 13.5

ขั้นตอนที่ 7 ทำการหาค่าสัมบูรณ์ของผลต่างระหว่างค่าเฉลี่ยที่ได้จากขั้นตอนที่ 5 และ 6 จากตัวอย่างข้างต้น ได้ว่า |8.5 – 13.5| = 5

ขั้นตอนที่ 8 นำผลต่างที่ได้จากขั้นตอนที่ 7 มาเพิ่ม (กรณี Sum′ > Sum) หรือลบออก (กรณี Sum′ < Sum) ให้กับชุดข้อมูลที่ผ่านการดึงค่าขอบเขตออกมาแล้ว จากตัวอย่างข้างต้น ได้ว่า {7 + 5, 8 + 5, 9 + 5, 10 + 5} = {12, 13, 14, 15}

ขั้นตอนที่ 9 นำค่าขอบเขตที่ดึงออกในขั้นตอนที่ 1 ใส่กลับเข้าในชุดข้อมูลที่ได้ในขั้นตอนที่ 8 จากตัวอย่างข้างต้น ได้ว่าง {6, 12, 13, 14, 15}

ทำการทดสอบหาผลรวมจากชุดข้อมูลที่ได้ พบว่า Sum{6, 12, 13, 14, 15} = 60 ตามที่ต้องการ โดยสามารถรักษาค่าน้อยที่สุดไว้ (ค่า 6) ได้

3.2.6 การจัดการค่าว่าง (Null)

ในฐานข้อมูล ค่าว่างเป็นคำที่บอกให้ทราบว่าในตำแหน่งข้อมูลข้อมูลนั้นยังไม่มีการกำหนดค่า โดยทั่วไปแล้วในฐานข้อมูลต่างๆ ล้วนมีค่าว่างเป็นจำนวนที่แตกต่างกันออกไป

ถ้าหากในชุดข้อมูลมีค่าว่าง ก่อนเริ่มต้นกระบวนการการปรับเปลี่ยนข้อมูลทุกครั้งต้องทำการเติมค่าที่อยู่ภายในขอบเขตด้วยการสุ่มค่าข้อมูลทั้ง 2 วิธีที่นำเสนอ เพราะค่าว่างไม่สามารถนำมาใช้ในกระบวนการการปรับเปลี่ยนข้อมูลได้ เมื่อเติมข้อมูลจนชุดข้อมูลไม่มีค่าว่างแล้วจึงสามารถนำชุดข้อมูลใหม่นี้มาใช้ในการปรับเปลี่ยนข้อมูลได้

ตัวอย่างสถานการณ์ในการนำไปใช้ ในองค์กรหนึ่งพบว่าข้อมูลที่กำลังนำมาปรับเปลี่ยนเพื่อนำไปใช้ในฐานข้อมูลฮันนี่พอตมีค่าว่างเป็นจำนวนมากและไม่ต้องการให้มีค่าว่างเกิดขึ้นในข้อมูลใหม่ การใช้วิธีการจัดการค่าว่างในงานวิจัยนี้สามารถช่วยกำจัดค่าว่างที่มีอยู่โดยการแทนที่ค่าที่เหมาะสมเพื่อให้ข้อมูลมีความแนบเนียนและตรงตามความต้องการขององค์กร โดยในงานวิจัยนี้ได้เสนอการจัดการค่าว่าง 4 วิธี ดังนี้

3.2.6.1 ทำการกำจัดค่าว่างที่มีอยู่ทั้งหมดโดยการแทนที่ค่าว่างด้วยค่าข้อมูลที่เหมาะสมด้วยวิธีการสุ่มค่าข้อมูลโดยควบคุมขอบเขตหรือการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล

3.2.6.2 ทำการรักษาจำนวนของค่าว่างไว้ให้เหมือนต้นฉบับ แต่ทำการสลับตำแหน่งของค่าว่างอย่างสุ่ม

3.2.6.3 ทำการกำหนดค่าว่างเป็นจำนวนตามที่ผู้ทำการปรับเปลี่ยนระบุเข้ามาโดยต้องไม่เกินจำนวนข้อมูลทั้งหมด และสุ่มตำแหน่งของค่าว่าง

3.2.6.4 ทำการกำหนดค่าว่างโดยสุ่มจำนวนของค่าว่างภายในช่วงที่ผู้ทำการปรับเปลี่ยนข้อมูลกำหนดเพื่อให้การจำนวนของค่าว่างแต่ละครั้งมีความแตกต่างกันออกไป

ตารางที่ 3.2 แสดงการจัดการค่าว่างในวิธีต่างๆ ที่นำเสนอในงานวิจัยนี้

แถวที่	ข้อมูลต้นฉบับ	ข้อมูลหลังการปรับเปลี่ยน	วิธีที่ 3.3.5.1	วิธีที่ 3.3.5.2	วิธีที่ 3.3.5.3	วิธีที่ 3.3.5.4
1	21	30	30	30	30	NULL
2	42	41	41	41	NULL	NULL
3	NULL	28	28	NULL	28	28
4	48	40	40	40	40	40
5	NULL	43	43	43	43	43
6	NULL	50	50	50	50	NULL
7	82	88	88	88	88	88
8	83	30	30	NULL	30	NULL
9	NULL	31	31	NULL	NULL	31
10	88	25	25	NULL	NULL	NULL

จากตารางที่ 3.2 เป็นตัวอย่างข้อมูลที่ได้จากปรับแต่งจำนวนของค่าว่าง โดยที่ข้อมูลต้นฉบับที่จำนวนค่าว่าง 4 ค่าในแถวที่ 3 5 6 และ 9 จากที่ได้กล่าวไว้ข้างต้น ก่อนทำการปรับเปลี่ยนข้อมูล ต้องทำการดึงค่าว่างออกแล้วนำข้อมูลที่เหลือมาผ่านกระบวนการการปรับเปลี่ยนข้อมูล ตัวอย่างของข้อมูลที่ได้แสดงในคอลัมน์ข้อมูลหลังการปรับเปลี่ยน ในการจัดการค่าว่างด้วยวิธีที่ 3.2.6.1 เป็นการแทนที่ค่าว่างด้วยค่าที่เหมาะสมที่ได้จากข้อมูลที่มีอยู่นำไปผ่านกระบวนการการปรับเปลี่ยนข้อมูล ในการจัดการค่าว่างด้วยวิธีที่ 3.2.6.2 เป็นการคงไว้ซึ่งจำนวนของค่าว่างแต่

ทำการสลับตำแหน่งของค่าว่างให้มีตำแหน่งแตกต่างไปจากข้อมูลต้นฉบับ การจัดการค่าว่างด้วยวิธีที่ 3.2.6.3 จากตัวอย่างในตารางทำการกำหนดให้มีค่าว่าง 3 จำนวนจึงทำการสร้างข้อมูลให้ครบก่อนแล้วทำการใส่ค่าว่างเข้าไปในภายหลัง และการจัดการค่าว่างด้วยวิธีที่ 3.2.6.4 คล้ายกับวิธีที่ 3.2.6.3 แต่เป็นการระบุเข้ามาเป็นช่วงเพื่อให้ระบบสุ่มจำนวนของค่าว่าง โดยในตัวอย่างเป็นการระบุช่วงจำนวนของค่าว่างเป็น 3 – 5 และระบบสุ่มจำนวนค่าว่างได้เป็น 5 จำนวน

3.2.7 การจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม

สำหรับข้อมูลตัวเลขทศนิยมนั้น หลังจากที่ผ่านกระบวนการการสุ่มข้อมูลแล้วได้ข้อมูลที่มีจำนวนตัวเลขที่แสดงหลังจุดทศนิยมแตกต่างไปจากข้อมูลต้นฉบับ หรือผู้ทำการปรับเปลี่ยนข้อมูลต้องการกำหนดจำนวนตัวเลขที่แสดงหลังจุดทศนิยมเพื่อให้ข้อมูลที่ได้มีลักษณะที่คล้ายคลึงหรือแตกต่างกับข้อมูลต้นฉบับ เพื่อตอบสนองความต้องการดังกล่าว

ตัวอย่างสถานการณ์ในการนำไปใช้ ในองค์กรหนึ่งพบว่าเมื่อผ่านกระบวนการการสุ่มค่าข้อมูลแล้ว ข้อมูลจำนวนจริงบางค่ามีตัวเลขที่แสดงหลังจุดทศนิยมแตกต่างกันออกไป บ้างก็ 2 หลัก บ้างก็ 3 หลัก แต่ผู้ทำการปรับเปลี่ยนข้อมูลต้องการให้ข้อมูลทั้งหมดเป็นทศนิยม 2 หลัก จึงต้องนำข้อมูลผ่านกระบวนการการจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม โดยในงานวิจัยนี้ได้เสนอการจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม 3 วิธี ดังนี้

3.2.7.1 ทำการรักษาจำนวนตัวเลขที่แสดงหลังจุดทศนิยมไว้ให้เหมือนกับข้อมูลต้นฉบับ เช่น ข้อมูลต้นฉบับเป็น 4.55 เมื่อทำการสุ่มข้อมูลแล้วได้ข้อมูลใหม่เป็น 6.312 เมื่อเลือกการจัดการตัวเลขที่แสดงหลังจุดทศนิยมด้วยวิธีนี้ ผลลัพธ์ที่ได้คือ 6.31

3.2.7.2 ทำการกำหนดตัวเลขที่แสดงหลังจุดทศนิยมตามที่ผู้ทำการปรับเปลี่ยนข้อมูลระบุ โดยที่ข้อมูลทุกตัวจะมีจำนวนตัวเลขที่แสดงหลังจุดทศนิยมเหมือนกันทั้งชุดข้อมูล เช่น ข้อมูลต้นฉบับ {5.123, 6.881, 9.013} เมื่อกำหนดให้แสดงแค่ทศนิยม 2 ตำแหน่ง ผลลัพธ์ที่ได้คือ {5.12, 6.88, 9.01} เป็นต้น

3.2.7.3 ทำการกำหนดตัวเลขที่แสดงหลังจุดทศนิยมโดยการสุ่มค่าตัวเลขภายในช่วงที่ผู้ทำการปรับเปลี่ยนข้อมูลกำหนดขึ้นมา เพื่อให้เกิดความหลากหลายของตัวเลขที่แสดงหลังจุดทศนิยม เช่น กำหนดช่วง 1-4 หมายความว่า ข้อมูลสามารถมีตัวเลขทศนิยมได้ 1 ถึง 4 ตำแหน่ง ตัวอย่างเช่น {5.12, 6.123, 7.9, 10.11} เป็นต้น

3.2.8 การจัดการข้อมูลที่มีรูปแบบตรงกัน

เป็นการจัดการให้ข้อมูลใหม่ที่ได้หลังจากการสุ่มค่าข้อมูลมีรูปแบบที่ตรงกับข้อมูลต้นฉบับในกรณีที่ข้อมูลต้นฉบับเป็นข้อมูลที่มีรูปแบบ เช่น ในหลักหน่วยของข้อมูลทุกค่าเป็นเลข 0 เป็นต้น

ตัวอย่างสถานการณ์ในการนำไปใช้ ข้อมูลเงินเดือนซึ่งเป็นข้อมูลประเภทหนึ่งที่หลายองค์กรมีไว้เพื่อจัดเก็บข้อมูลเงินเดือนของบุคลากรภายในองค์กร ส่วนใหญ่มีลักษณะที่เป็นแบบแผน เช่น บางองค์กรมีลักษณะข้อมูลเงินเดือนโดยหลักหน่วยลงท้ายด้วยเลข 0 เช่น {25650, 35550, 55320} บาท เป็นต้น หรือบางองค์กรลงท้ายทั้งหลักหน่วยและหลักสิบด้วยเลข 0 เช่น {25600, 35500, 55300} บาท เป็นต้น ดังนั้นหลังจากที่ข้อมูลผ่านกระบวนการการสุ่มค่าข้อมูลและปรับค่าขอบเขตแล้ว การจัดการรูปแบบของชุดข้อมูลให้อยู่ในรูปแบบตามข้อมูลต้นฉบับจึงเป็นอีกวิธีหนึ่งที่ช่วยให้ข้อมูลที่ได้มีความแนบเนียนมากยิ่งขึ้น ขั้นตอนในการจัดการข้อมูลที่มีรูปแบบตรงกัน มีดังนี้

ขั้นตอนที่ 1 ทำการปรับความยาวของข้อมูลให้มีจำนวนหลักเท่ากัน เนื่องจากในบางกรณีที่ข้อมูลชุดเดียวกันมีทั้งจำนวนในหลักพัน หลักหมื่น หรือหลักแสน เพื่อความง่ายในการจัดการข้อมูลจึงต้องทำการปรับให้มีจำนวนหลักเท่ากันโดยการเพิ่มเลข 0 ไว้ข้างหน้าโดยยึดความยาวของข้อมูลที่มีจำนวนหลักของตัวเลขมากที่สุด (ที่ไม่ใช่ตัวเลขทศนิยม) เช่น ในชุดข้อมูลประกอบด้วย {5000, 45000, 600000} เนื่องจาก 600000 เป็นข้อมูลที่มีจำนวนหลักมากที่สุด (6 หลัก) จึงทำการเพิ่มจำนวนหลักให้กับ 5000 และ 45000 โดยการเติม 0 ไว้หน้าข้อมูล จะได้เป็น 005000 และ 045000 ตามลำดับ

ขั้นตอนที่ 2 ทำการหารูปแบบของชุดข้อมูล โดยการพิจารณาข้อมูลแต่ละหลักของข้อมูลแต่ละตัว ถ้าหากข้อมูลในหลักนั้นเป็นตัวเลขเดียวกันทั้งชุดข้อมูล โดยในงานวิจัยนี้พิจารณาเฉพาะเลข 0 เท่านั้น ดังนั้นถ้าหลักใดเป็นเลข 0 ทั้งชุดข้อมูล ให้กำหนดรูปแบบเป็น “0” แต่ถ้าหากในหลักนั้นมีข้อมูลบางตัวที่เป็นเลขอื่นที่ไม่ใช่ 0 อยู่ด้วย ให้กำหนดรูปแบบเป็น “X” ตัวอย่างเช่น ในชุดข้อมูลประกอบด้วย {005000, 045000, 600000} เมื่อพิจารณาหลักหน่วย พบว่าข้อมูลทุกตัวมีหลักหน่วยเป็นเลข 0 ดังนั้น จึงมีรูปแบบเป็น “0” ต่อมาพิจารณาหลักสิบ พบว่าข้อมูลทุกตัวมีหลักสิบเป็นเลข 0 ดังนั้น จึงมีรูปแบบเป็น “00” หมายความว่า ทั้งหลักสิบและหลักหน่วยต้องเป็นเลข 0 ต่อมาพิจารณาหลักร้อย พบว่าข้อมูลทุกตัวมีหลักร้อยเป็นเลข 0 ดังนั้น จึงมีรูปแบบเป็น “000” ต่อมาพิจารณาหลักพัน พบว่าข้อมูลทุกตัวมีหลักพันที่แตกต่างกัน นั่นคือ ประกอบด้วย {5, 5, 0} ดังนั้นรูปแบบที่ได้คือ “X000” ทำตามขั้นตอนนี้จนกระทั่งครบทุกหลักของข้อมูล ซึ่งจากตัวอย่าง รูปแบบที่ได้คือ “XXX000”

ขั้นตอนที่ 3 ทำการหาว่า “X” ตัวสุดท้ายในรูปแบบที่ได้จากขั้นตอนที่ 2 อยู่ตำแหน่งใดเพื่อทำการตัด “0” ทุกตัวที่อยู่หลัง “X” ตัวสุดท้ายออกไป เพราะว่าข้อมูลในหลักที่มีรูปแบบเป็น “0” จะต้องถูกแทนที่กลับด้วย “0” ในขั้นตอนท้าย จึงไม่มีความจำเป็นที่จะต้องนำมาทำการสุ่มค่าข้อมูล ข้อมูลที่ควรนำมาพิจารณาในการสุ่มค่าคือข้อมูลที่มีรูปแบบเป็น “X” เมื่อได้ตำแหน่ง “X” ตัวสุดท้ายในรูปแบบแล้ว กลับมาที่ข้อมูลในชุดข้อมูล ให้ทำการตัดข้อมูลตั้งแต่ข้อมูลหลังตำแหน่ง “X” ตัวสุดท้ายในรูปแบบ ตัวอย่างเช่น ข้อมูล {52000, 45000, 60000} มีรูปแบบเป็น “XX000” ทำการตัดข้อมูลตั้งแต่หลังตำแหน่ง “X” ตัวสุดท้าย จากตัวอย่าง “X” ตัวสุดท้ายอยู่ในตำแหน่งที่ 2 จึงทำการตัดข้อมูลตั้งแต่ตำแหน่งที่ 3 เป็นต้นไป ทำให้ได้ข้อมูลใหม่เป็น {52, 45, 60} และเมื่อตัดข้อมูลเสร็จสิ้น ให้จัดเก็บรูปแบบที่ถูกตัดไปด้วย จากในตัวอย่าง รูปแบบที่ถูกตัดไปคือ “000” ต้องถูกจัดเก็บไว้เพื่อนำไปใส่กลับเข้าไปยังข้อมูลในขั้นตอนท้าย

ขั้นตอนที่ 4 นำข้อมูลที่ผ่านการตัดแล้วเข้ากระบวนการสุ่มค่าข้อมูลหรือการปรับเปลี่ยนค่าขอบเขตดังที่ได้นำเสนอไว้ในงานวิจัย

ขั้นตอนที่ 5 ทำการนำรูปแบบที่ตัดออกไปในขั้นตอนที่ 3 ใส่กลับเข้าไปยังข้อมูลที่ได้จากขั้นตอนที่ 4 จากตัวอย่างในขั้นตอนที่ 3 {52, 45, 60} เมื่อผ่านกระบวนการตามขั้นตอนที่ 4 แล้วได้ข้อมูลใหม่เป็น {50, 59, 47} จากนั้นนำรูปแบบที่ถูกตัดออกไป ซึ่งในที่นี้คือ “000” กลับมาใส่ในตำแหน่งเดิมให้กับข้อมูลทุกตัว ได้เป็น {50000, 59000, 47000}

ขั้นตอนที่ 6 กรณีที่รูปแบบมี “0” อยู่ระหว่าง “X” เช่น “X0X00” ให้ทำการแทนที่ “0” ลงไปในข้อมูลในหลักนั้นๆ เพื่อให้ข้อมูลมีความแนบเนียนมากยิ่งขึ้น ตัวอย่างเช่น ข้อมูล {10300, 20400, 50600} มีรูปแบบเป็น “X0X00” นำไปผ่านขั้นตอนที่ 3 ได้ข้อมูลเป็น {103, 204, 506} จากนั้นนำไปผ่านกระบวนการการสุ่มค่าข้อมูล ได้ข้อมูลตัวอย่างเป็น {425, 321, 221} จากนั้นทำขั้นตอนที่ 5 ได้ข้อมูลเป็น {42500, 32100, 22100} และขั้นตอนสุดท้าย ทำการแทนที่รูปแบบ “0” ที่อยู่ระหว่างรูปแบบ “X” ในข้อมูลทุกตัว ได้ข้อมูลเป็น {40500, 30100, 20100}

เมื่อทำครบทุกขั้นตอน ข้อมูลที่ได้ยังคงมีรูปแบบตามข้อมูลต้นฉบับและมีคุณสมบัติตามจุดประสงค์ของงานวิจัยฉบับนี้ บางกรณีที่ผู้ทำการปรับเปลี่ยนไม่ต้องการให้ข้อมูลรักษารูปแบบเดิมไว้ กระบวนการนี้จึงอยู่ที่การตัดสินใจของผู้ทำการปรับเปลี่ยนข้อมูลว่าต้องการหรือไม่ต้องการให้มีการรักษารูปแบบของข้อมูล

3.2.9 **การใช้หลายวิธีร่วมกัน**

ในแต่ละวิธีที่ได้นำเสนอในงานวิจัยนี้ล้วนให้ข้อมูล ลักษณะการกระจายของข้อมูล ที่ตำแหน่งการปรับเปลี่ยนของข้อมูล และลักษณะของข้อมูลที่แตกต่างกัน การใช้หลายวิธีร่วมกันจึงสามารถช่วยเพิ่มความแนบเนียนและความปลอดภัยให้กับข้อมูลต้นฉบับมากยิ่งขึ้น

ตัวอย่างสถานการณ์ในการนำไปใช้ การสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล ดังที่ได้กล่าวไว้ข้างต้น ข้อมูลทุกค่าเกิดการเปลี่ยนแปลงยกเว้นค่ามากที่สุดและน้อยที่สุด การนำวิธีการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดมาใช้ร่วมด้วย ทำให้ข้อมูลทุกค่ารวมถึงค่ามากที่สุดและน้อยที่สุดมีการเปลี่ยนแปลง หรือการใช้การสุ่มค่าข้อมูลร่วมกับการปรับเปลี่ยนค่าผลรวมและการจัดการค่าว่าง ทำให้ข้อมูลที่ได้มีการเปลี่ยนแปลงทุกค่าและจำนวนและตำแหน่งของค่าว่างมีการเปลี่ยนแปลงแตกต่างไปจากข้อมูลต้นฉบับ เป็นต้น ดังนั้นการใช้หลายวิธีร่วมกันช่วยให้สามารถตอบสนองความต้องการของผู้ทำการปรับเปลี่ยนข้อมูลได้มากขึ้น รวมไปถึงข้อมูลที่ได้สามารถบรรลุวัตถุประสงค์ของการปรับเปลี่ยนข้อมูลได้ดียิ่งขึ้น

ในบทนี้ได้นำเสนอวิธีการปรับเปลี่ยนข้อมูลด้วยวิธีการ 1) การสุ่มค่าข้อมูลโดยควบคุมขอบเขต และ 2) การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล 3) การสลับที่ของข้อมูล 4) การแก้ไขค่ามากที่สุดและน้อยที่สุด การแก้ไขข้อมูลเชิงสถิติด้วยวิธีการ 5) การปรับเปลี่ยนค่าผลรวม การจัดการรูปแบบและลักษณะของข้อมูลด้วยวิธีการ 6) การจัดการค่าว่าง 7) การจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม และ 8) การจัดการข้อมูลที่มีรูปแบบตรงกัน และการจัดการข้อมูลเพื่อให้ข้อมูลมีความปลอดภัยมากยิ่งขึ้นด้วยวิธีการ และ 9) การใช้หลายวิธีร่วมกัน สามารถสรุปลักษณะข้อมูลที่ได้และลักษณะการใช้งานของแต่ละวิธีได้ในตารางที่ 3.3

ตารางที่ 3.3 สรุปลักษณะข้อมูลที่ได้และความลักษณะการใช้งานของแต่ละวิธี

วิธีการ	ลักษณะข้อมูลที่ได้	ลักษณะการใช้งาน
การสุ่มค่าข้อมูลโดยควบคุมขอบเขต	ข้อมูลทุกค่าเกิดการเปลี่ยนแปลงอย่างสุ่ม	ใช้ในกรณีที่ไม่สนใจลักษณะการกระจายของข้อมูล
การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล	ข้อมูลทุกค่าเกิดการเปลี่ยนแปลงอย่างสุ่ม	ใช้ในกรณีที่ต้องการให้ข้อมูลใหม่และข้อมูลต้นฉบับมีค่าใกล้เคียงกัน
การสลับที่ของข้อมูล	ข้อมูลมีการเปลี่ยนตำแหน่ง แต่ลักษณะข้อมูลเดิมยังคงอยู่	ใช้ในกรณีที่ต้องการให้ข้อมูลใหม่มีการกระจายเหมือนข้อมูลต้นฉบับ
การปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุด	ค่ามากที่สุดและน้อยที่สุดมีการเปลี่ยนแปลงตามที่ระบุเข้ามา	ใช้ในกรณีที่ต้องการแก้ไขค่าขอบเขต
การปรับเปลี่ยนค่าผลรวมกรณีเก็บรักษาค่าขอบเขต	ข้อมูลทุกค่ามีการเปลี่ยนแปลงขึ้นอยู่กับค่าผลรวมที่ระบุเข้ามา เฉพาะค่าขอบเขตที่ยังคงเป็นค่าเดิม	ใช้ในกรณีที่ต้องการแก้ไขค่าผลรวมโดยรักษาค่ามากที่สุดหรือน้อยที่สุดไว้
การปรับเปลี่ยนค่าผลรวมกรณีไม่เก็บรักษาค่าขอบเขต	ข้อมูลทุกค่ามีการเปลี่ยนแปลงขึ้นอยู่กับค่าผลรวมที่ระบุเข้ามา	ใช้ในกรณีที่ต้องการแก้ไขค่าผลรวมโดยไม่รักษาค่ามากที่สุดหรือน้อยที่สุดไว้
การจัดการค่าว่าง	จำนวนค่าว่ามีการเปลี่ยนแปลง	ใช้ในกรณีที่ต้องการแก้ไขจำนวนค่าว่างในชุดข้อมูล
การจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม	รูปแบบการแสดงจำนวนตัวเลขที่แสดงหลังจุดทศนิยมมีการเปลี่ยนแปลง	ใช้ในกรณีที่ต้องการแก้ไขการแสดงผลของจำนวนตัวเลขที่แสดงหลังจุดทศนิยม
การใช้หลายวิธีร่วมกัน	ขึ้นอยู่กับการปรับแต่ง	เพื่อให้ข้อมูลเป็นไปตามที่ต้องการมากยิ่งขึ้น

แต่ละวิธีที่นำเสนอในบทนี้ทำให้ได้ข้อมูลที่แตกต่างกันออกไป ในบทถัดไปเป็นการทดสอบว่าวิธีการที่นำเสนอในงานวิจัยนี้สามารถปรับเปลี่ยนข้อมูลให้ออกมาในลักษณะใดได้บ้าง

และมีมาตรอะไรบ้างในการประเมินแต่ละขั้นตอนว่าสามารถบรรลุวัตถุประสงค์ของงานวิจัยได้มากน้อยเพียงใด

บทที่ 4

การทดสอบและผลการทดสอบ

ในบทนี้เป็นการทดสอบขั้นตอนวิธีที่เสนอในงานวิจัยที่ได้กล่าวไว้ในบทที่ 3 ด้วยวิธีทางสถิติต่างๆ เพื่อพิสูจน์ว่าวิธีที่นำเสนอนั้นสามารถบรรลุวัตถุประสงค์ของงานวิจัยได้

4.1 ขั้นตอนการทดสอบ

หลังจากการที่ข้อมูลต้นฉบับผ่านกระบวนการการปรับเปลี่ยนข้อมูลในขั้นตอนต่างๆ ที่ได้นำเสนอในงานวิจัยนี้แล้ว จากนั้นทำการนำข้อมูลที่ได้มาทดสอบด้วยวิธีทางสถิติโดยอาศัยวัตถุประสงค์ของงานวิจัยเป็นเป้าหมายหลัก ประกอบด้วย 2 ส่วนหลัก ได้แก่

4.1.1 การทดสอบข้อมูลด้วยการคำนวณหาค่าสถิติ เป็นการนำข้อมูลที่ได้มาคำนวณด้วยค่าสถิติที่เหมาะสมเพื่อให้สามารถมองเห็นลักษณะของข้อมูลได้และง่ายต่อการเปรียบเทียบระหว่างข้อมูลต้นฉบับและข้อมูลหลังการปรับเปลี่ยน ประกอบด้วย

4.1.1.1 การทดสอบความแนบเนียนของข้อมูลโดยอาศัยวิธีการหาร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้

4.1.1.2 การทดสอบความสามารถในการปกปิดข้อมูลต้นฉบับ โดยอาศัยวิธีการหาค่าความผิดพลาดในการปกปิดข้อมูลต้นฉบับ

4.1.1.3 การทดสอบความแตกต่างในเชิงสถิติ โดยอาศัยวิธีการหาความสัมพันธ์ระหว่างข้อมูล 2 ตัวแปร

4.1.2 การทดสอบข้อมูลด้วยการพิจารณาจากกราฟข้อมูล เป็นการนำข้อมูลมาแสดงให้อยู่ในรูปของกราฟแท่งเพื่อเปรียบเทียบการกระจายของข้อมูลระหว่างข้อมูลต้นฉบับและข้อมูลหลังการปรับเปลี่ยน

4.2 เครื่องมือที่ใช้ในการวิจัย

ในส่วนของเครื่องมือวิจัยแบ่งออกเป็น 2 ส่วนคือ ส่วนอุปกรณ์หมายถึงเครื่องคอมพิวเตอร์ที่ใช้ในการวิจัย และส่วนชุดคำสั่ง หมายถึงส่วนประกอบทั้งหมดที่ถูกนำมาใช้ร่วมกันในการสร้างโปรแกรมขึ้นสำหรับงานวิจัยนี้

4.2.1 ส่วนอุปกรณ์ (Hardware) ประกอบด้วย

1) ระบบปฏิบัติการวินโดวส์เซเว่น (Windows 7)

2) ซีพียู อินเทล (Intel Core 2 Duo 2.10Ghz)

3) หน่วยความจำ 4 กิกะไบต์

4.2.2 ส่วนชุดคำสั่ง (Software) ประกอบด้วย

4.2.2.1 ภาษาการเขียนโปรแกรม (Programming Language)

1) ภาษาซีชาร์พ เป็นภาษาหลักในการใช้สร้างโปรแกรม

2) ภาษาเอสคิวแอล เป็นภาษาสำหรับการสอบถามข้อมูล

4.2.2.2 ฐานข้อมูล (Database Application)

1) MS SQL Server 2008 R2 เป็นโปรแกรมจัดการฐานข้อมูล

4.3 เกณฑ์ในการวัดประสิทธิภาพของขั้นตอนวิธี

4.3.1 ค่าความผิดพลาดในการปิดบัง (Hiding Failure หรือ HF)

Oliveira และ Zaiane [28] ได้เสนอมาตรในการวัดความสามารถในการปกปิดข้อมูลที่องค์กรต้องการป้องกัน (Sensitive Information) โดยได้อธิบายไว้ว่าค่าความผิดพลาดในการปิดบังเป็นค่าสัดส่วนของข้อมูลที่ต้องการป้องกันที่ไม่ถูกปกปิดด้วยขั้นตอนวิธีในการปรับเปลี่ยนข้อมูล ซึ่งค่าความผิดพลาดในการปิดบังที่ดีที่สุดนั้นควรมีค่าเป็น 0 นั่นคือไม่มีข้อมูลต้นแบบปรากฏในข้อมูลที่ผ่านการปรับเปลี่ยน สูตรที่ใช้ในการคำนวณหาค่าความผิดพลาดในการปิดบัง คือ

$$\text{HF} = \frac{\#R_p(D')}{\#R_p(D)}$$

โดยกำหนด

R_p ย่อมาจาก Restricted Pattern

D' คือ ชุดข้อมูลที่ผ่านการปรับเปลี่ยน

D คือ ชุดข้อมูลต้นแบบ

$\#R_p(D')$ คือ จำนวนของรูปแบบที่ต้องถูกปกปิดปรากฏในข้อมูลที่ถูกปรับเปลี่ยน

$\#R_p(D)$ คือ จำนวนของรูปแบบที่ต้องถูกปกปิดปรากฏในข้อมูลต้นฉบับ

ในที่นี้กำหนดให้ข้อมูลต้นฉบับแต่ละแถวในฐานข้อมูลเป็นข้อมูลที่ต้องถูกปกปิด ทำให้ $\#R_p$ (D) มีค่าเท่ากับจำนวนข้อมูลต้นฉบับ และ $\#R_p$ (D′) เป็นจำนวนของข้อมูลที่ถูกปรับเปลี่ยนตรงกับค่าของข้อมูลต้นฉบับในแถวเดียวกัน ตัวอย่างดังตารางที่ 4.1

ตารางที่ 4.1 ตัวอย่างข้อมูลและผลที่ได้ในการหาค่าความผิดพลาดในการปิดบัง

แถวที่	**ข้อมูลต้นฉบับ**	**ข้อมูลที่ถูกปรับเปลี่ยน**	R_p (D′)
1	5	5	1
2	10	6	0
3	11	4	0
4	8	7	0
5	7	7	1
		$\#R_p$ (D′) =	2
		HF =	2/5 = 0.4

จากข้อมูลในตารางที่ 4.1 เมื่อมาคำนวณหาค่า HF จะได้ว่า HF = 0.4 หมายความว่า ยังคงมีข้อมูลที่ผ่านการปรับเปลี่ยนบางแถวที่มีค่าตรงกับข้อมูลต้นฉบับ ยิ่งค่า HF มากขึ้นเท่าไหร่ ยิ่งมีจำนวนข้อมูลที่ผ่านการปรับเปลี่ยนที่ตรงกับข้อมูลต้นฉบับในแถวเดียวกันเพิ่มขึ้นมากเท่านั้น ส่งผลให้โอกาสที่ข้อมูลต้นฉบับเกิดการรั่วไหลมากยิ่งขึ้น ในทางตรงกันข้าม ยิ่งค่า HF น้อยลงเท่าไหร่ ยิ่งมีจำนวนข้อมูลที่ผ่านการปรับเปลี่ยนที่ตรงกับข้อมูลต้นฉบับในแถวเดียวกันน้อยลงเท่านั้น และโอกาสที่ข้อมูลต้นฉบับเกิดการรั่วไหลยิ่งน้อยลงไป และเมื่อค่า HF = 0 แสดงว่าข้อมูลทุกค่าไม่มีแถวใดที่ตรงกับข้อมูลต้นฉบับ และข้อมูลต้นฉบับได้รับการปกปิดไว้อย่างสมบูรณ์

4.3.2 ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้

เนื่องจากในงานวิจัยนี้กำหนดให้ขอบเขตของข้อมูลที่เป็นไปได้ซึ่งหมายถึงช่วงข้อมูลระหว่างค่ามากที่สุดและน้อยที่สุดของข้อมูลต้นฉบับ ข้อมูลที่อยู่ในช่วงดังกล่าวเป็นข้อมูลที่มีความแนบเนียนเพียงพอที่ทำให้ผู้โจมตีข้อมูลไม่สามารถตระหนักได้ว่าตนเองกำลังมีปฏิสัมพันธ์กับข้อมูลปลอมอยู่ ดังนั้นในการหาสัดส่วนระหว่างจำนวนข้อมูลที่อยู่ในขอบเขตข้อมูลที่เป็นไปได้นั้นและจำนวนข้อมูลทั้งหมด สามารถช่วยให้พิจารณาได้ว่าขั้นตอนวิธีที่นำเสนอในงานวิจัยนี้สามารถก่อให้เกิดข้อมูลที่มีความแนบเนียนได้ โดยสมการที่ใช้ในการหาสัดส่วน ดังนี้

$$\text{ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้} = \frac{\#Inbound}{N} \text{ X } 100$$

โดยที่

#Inbound คือ จำนวนของข้อมูลหลังการปรับเปลี่ยนที่อยู่ในขอบเขต

N คือ จำนวนของข้อมูลต้นฉบับ

ยิ่งค่าร้อยละของข้อมูลที่อยู่ในของเขตข้อมูลที่เป็นไปได้มากเท่าไหร่แสดงว่าข้อมูลที่ได้หลังการปรับเปลี่ยนมีความเชื่อถือมากขึ้นเท่านั้น สำหรับกรณีของการแก้ไขค่าขอบเขตและค่าผลรวม ค่าร้อยละของข้อมูลที่อยู่ภายในขอบเขตข้อมูลมีความเปลี่ยนแปลงแตกต่างกันออกไปขึ้นอยู่กับปรับค่าแก้ไขของผู้ทำการปรับตั้งค่า ตัวอย่างดังตารางที่ 4.2

ตารางที่ 4.2 ตัวอย่างข้อมูลและผลที่ได้จากการหาร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้

แถวที่	**ข้อมูลต้นฉบับ**	**ข้อมูลที่ถูกปรับเปลี่ยน**	**#Inbound (D')**
1	5	4	0
2	6	8	1
3	7	9	1
4	8	15	0
5	9	6	1
ค่ามากที่สุด = 9 ค่าน้อยที่สุด = 5		#Inbound (D') =	3
ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้ =			(2 * 100)/5 = 60%

4.3.3 **การหาความสัมพันธ์โดยการหาค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน** (Pearson Correlation Coefficient)

ในหลักสถิติมีวิธีการทางสถิติหลายวิธีที่ใช้ในการตรวจสอบและทดสอบความสัมพันธ์ระหว่างตัวแปร กรณีตัวแปรทั้ง 2 สามารถนำมาคำนวณได้ แต่ที่นิยมใช้กันทั่วไปคือการหาค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน หรือสหสัมพันธ์อย่างง่าย (Simple Correlation) โดยใช้สัญลักษณ์ r โดยมีสูตรที่ใช้ในการคำนวณดังนี้

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

โดยกำหนด

X, Y คือ ค่าที่ ของข้อมูลชุดที่ 1 และ 2

n คือ จำนวนข้อมูลของแต่ละชุดข้อมูล ซึ่งต้องมีจำนวนเท่ากัน

การตีความหมายความสัมพันธ์จากค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน สามารถตีความโดยพิจารณาจากตารางที่ 4.3

ตารางที่ 4.3 การตีความหมายความสัมพันธ์จากค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน

ค่า r	**ความสัมพันธ์ระหว่างข้อมูล**
r = .50 ถึง 1.00 หรือ r = -.50 ถึง -1.00	ถือว่าข้อมูลมีความสัมพันธ์ในระดับสูง
r = .30 ถึง .49 หรือ r = -.30 ถึง -.49	ถือว่าข้อมูลมีความสัมพันธ์ในระดับปานกลาง
r = .10 ถึง .29 หรือ r = -.10 ถึง -.29	ถือว่าข้อมูลมีความสัมพันธ์ในระดับต่ำ
r = .00	ถือว่าข้อมูลไม่มีความสัมพันธ์กัน

ในการหาลักษณะความสัมพันธ์ระหว่างตัวแปรนั้นเราสามารถสร้างแผนภาพการกระจาย (Scatter Diagram) เพื่อดูทิศทางของความสัมพันธ์ได้ โดยมีลักษณะความสัมพันธ์ 3 แบบ คือ

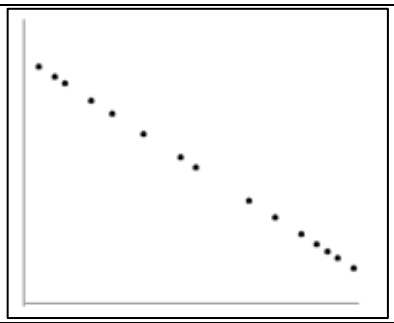
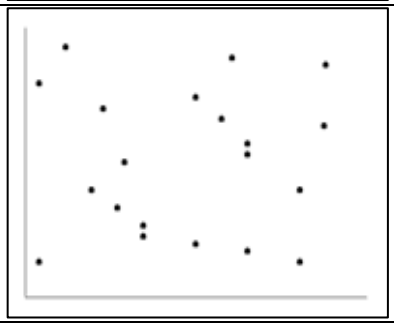
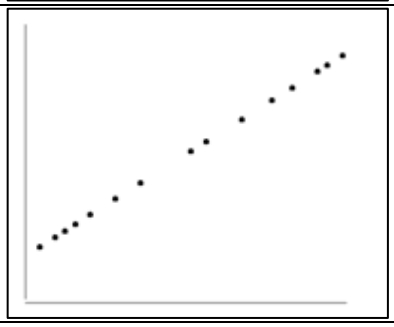
1.สหสัมพันธ์ทางบวก (Positive Correlations) ซึ่งหมายความว่าเมื่อตัวแปรตัวหนึ่งเพิ่มหรือลดลงอีกตัวแปรหนึ่งก็จะเพิ่มขึ้นหรือลดลงไปด้วย

2. สหสัมพันธ์ทางลบ (Negative Correlations) หมายถึงเมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มขึ้นหรือลดลงอีกตัวหนึ่งจะมีค่าเพิ่มหรือลดลงตรงข้ามเสมอ

3. สหสัมพันธ์เป็นศูนย์ (Zero Correlations) หมายถึงตัวแปรสองตัวไม่มีความสัมพันธ์ซึ่งกันและกัน

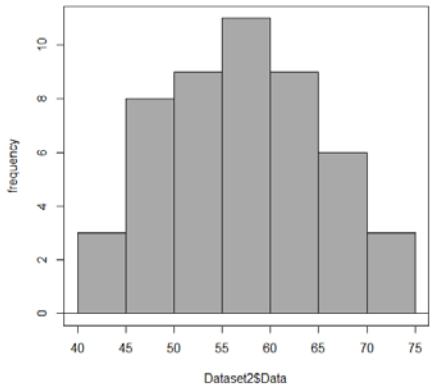
ลักษณะของแผนภาพการกระจายและความสัมพันธ์ของข้อมูลแสดงในตารางที่ 4.4

ตารางที่ 4.4 ลักษณะของแผนภาพการกระจายและความสัมพันธ์ของข้อมูลสำหรับค่า r ต่างๆ

ค่า r	แผนภาพการกระจาย	ความสัมพันธ์ของข้อมูล
r = -1		ความสัมพันธ์ทางลบ กล่าวคือ ถ้าตัวแปรหนึ่งมีค่าเพิ่ม (หรือลด) อีกตัวแปรหนึ่งจะมีค่าลด (หรือเพิ่ม)
r = 0		ไม่มีความสัมพันธ์แบบเชิงเส้นตรงต่อกัน
r = 1		ความสัมพันธ์ทางบวก กล่าวคือ ถ้าตัวแปรหนึ่งมีค่าเพิ่ม (หรือลด) อีกตัวแปรหนึ่งจะมีค่าเพิ่มขึ้น (หรือลด) ตาม

4.3.4 การกระจายของข้อมูล

การพิจารณาการกระจายของข้อมูลในงานวิจัยนี้ใช้ฮิสโทแกรม (Histogram) เป็นการนำเสนอข้อมูลด้วยการใช้กราฟแท่งแบบเฉพาะ โดยที่แกนตั้งแทนข้อมูลของความถี่ของข้อมูลและแกนนอนแทนข้อมูลของข้อมูลของคุณสมบัติที่เรากำลังสนใจ การนำฮิสโทแกรมมาใช้ตรวจสอบการกระจายของข้อมูลสามารถใช้ในการตรวจสอบความผิดปกติของข้อมูล เปรียบเทียบข้อมูลกับเกณฑ์ที่กำหนด หรือวิเคราะห์ข้อมูลทางสถิติ เป็นต้น ในงานวิจัยนี้พิจารณาเพียงความแตกต่างของการกระจายของข้อมูล ตัวอย่างของฮิสโทแกรมแสดงในรูปที่ 4.1



รูปที่ 4.1 ตัวอย่างการแสดงข้อมูลด้วยฮิสโทแกรม

4.4 ผลการทดสอบงานวิจัย

เนื่องจากงานวิจัยนี้ได้ใช้วิธีการสุ่มค่าเป็นวิธีหลักในการปรับเปลี่ยนข้อมูล ดังนั้นในแต่ละครั้งของการปรับเปลี่ยนข้อมูล ข้อมูลที่ได้มีความแตกต่างกันออกไปขึ้นอยู่กับลักษณะของข้อมูล ช่วงของขอบเขตข้อมูลที่เป็นไปได้ จำนวนข้อมูล เป็นต้น ข้อมูลที่ใช้ทดสอบประกอบด้วย

1) ข้อมูลตัวอย่าง เป็นข้อมูลจำนวนเต็มบวก ประกอบด้วยข้อมูลจำนวน 20 แถว มีค่ามากที่สุด = 34 และ ค่าน้อยที่สุด = 20 ค่าผลรวม = 517 มีการกระจายแบบปกติ
2) ข้อมูลอายุของบุคลากรในองค์กรหนึ่ง เป็นข้อมูลจำนวนเต็มบวก ประกอบด้วยข้อมูลจำนวน 10,000 แถว มีค่ามากที่สุด = 18 และ ค่าน้อยที่สุด = 36 ค่าผลรวม = 250417 มีการกระจายแบบปกติ
3) ข้อมูลเกรดเฉลี่ยของสถานศึกษาแห่งหนึ่ง เป็นข้อมูลตัวเลขทศนิยม ประกอบด้วยข้อมูลจำนวน 10,000 แถว มีค่ามากที่สุด = 4.00 และ ค่าน้อยที่สุด = 0.58 ค่าผลรวม = 27813.91

4.4.1 การสุ่มค่าข้อมูลโดยควบคุมขอบเขต

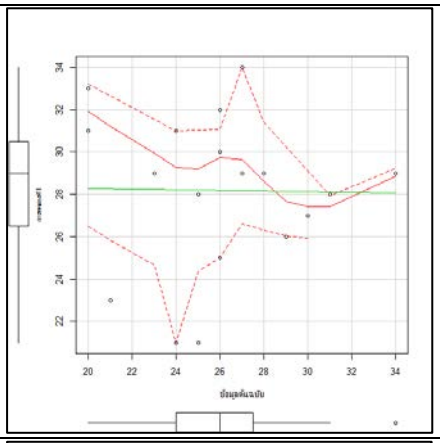
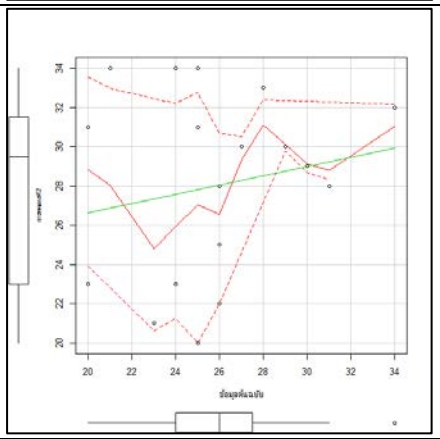
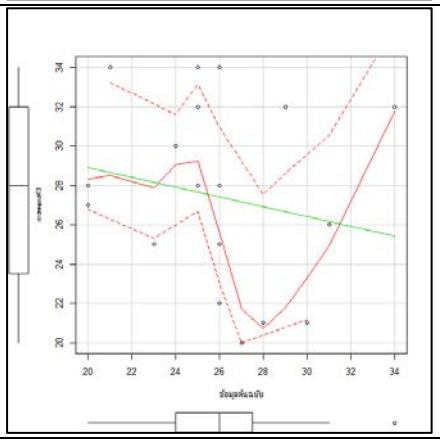
จากที่ได้กล่าวไว้ในบทที่ 3 การสุ่มค่าข้อมูลโดยควบคุมขอบเขตเป็นการปรับเปลี่ยนข้อมูลโดยการสุ่มค่าข้อมูลขึ้นมาใหม่โดยอาศัยค่ามากที่สุดและน้อยที่สุดที่ได้จากข้อมูลต้นฉบับเป็นขอบเขตของข้อมูลที่เป็นไปได้ การทดสอบการสุ่มค่าข้อมูลโดยควบคุมขอบเขตนี้อาศัยข้อมูลจากข้อมูลตัวอย่าง โดยทำการทดสอบโดยการสุ่มข้อมูลจำนวน 3 ครั้งแบ่งเป็นการทดสอบที่ 1 2 และ 3 ตามลำดับ ตัวอย่างผลการทดสอบที่ได้แสดงในตารางที่ 4.5

ตารางที่ 4.5 ตัวอย่างของข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขต

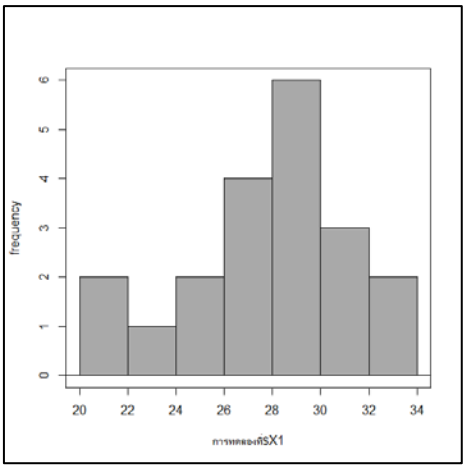
แถวที่	**ข้อมูลต้นฉบับ**	**ผลการทดสอบ 1**	**ผลการทดสอบ 2**	**ผลการทดสอบ 3**
1	24	21	34	30
2	29	26	30	32
3	26	30	25	25
4	20	33	23	28
5	26	25	22	34
6	24	31	23	30
7	25	21	31	32
8	26	30	28	28
9	21	23	34	34
10	31	28	28	26
11	30	27	29	21
12	26	32	22	22
13	28	29	33	21
14	23	29	21	25
15	27	29	30	20
16	25	28	34	34
17	34	29	32	32
18	27	34	30	20
19	25	28	20	28
20	20	31	31	27
ค่ามากที่สุด	**34**	**34**	**34**	**34**
ค่าน้อยที่สุด	**20**	**21**	**20**	**20**
ค่าความผิดพลาดในการปิดบัง		0	0	0
ร้อยละข้อมูลในขอบเขตข้อมูลที่เป็นไปได้		100	100	100
ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน		-0.0142	0.1768	-0.1812

จากตารางที่ 4.5 ข้อมูลที่ผ่านการปรับเปลี่ยนในแต่ละครั้งของการทดสอบสามารถปกปิดข้อมูลต้นฉบับได้เป็นอย่างดีเนื่องจากค่าความผิดพลาดในการปิดบังเป็นค่า 0 ทุกครั้งของการสุ่มค่าข้อมูล นอกจากนี้ในการปรับเปลี่ยนข้อมูลด้วยวิธีนี้ ข้อมูลทุกค่าอยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้เนื่องจากมีการควบคุมขอบเขตของข้อมูลไว้เพื่อป้องกันข้อมูลที่สุ่มขึ้นมาอยู่นอกขอบเขตของข้อมูลที่เป็นไปได้ และเมื่อพิจารณาความสัมพันธ์ระหว่างข้อมูลต้นฉบับด้วยค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน พบว่า ในแต่ละครั้งของการทดสอบข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยนแล้วมีความสัมพันธ์กันน้อยมากจนถึงไม่มีความสัมพันธ์เลย โดยพิจารณาจากแผนภาพการกระจายจากตารางที่ 4.6

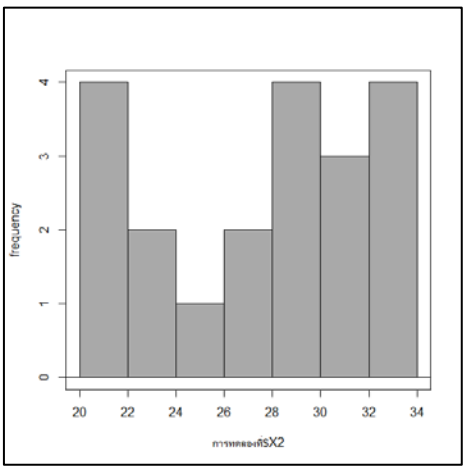
ตารางที่ 4.6 แผนภาพการกระจายของตัวอย่างข้อมูลในตารางที่ 4.5

ผลการทดสอบที่	แผนภาพการกระจาย	ความสัมพันธ์ระหว่างชุดข้อมูล
1		ข้อมูลมีความสัมพันธ์ระดับต่ำ
2		ข้อมูลไม่มีความสัมพันธ์กัน
3		ข้อมูลมีความสัมพันธ์กันปานกลาง

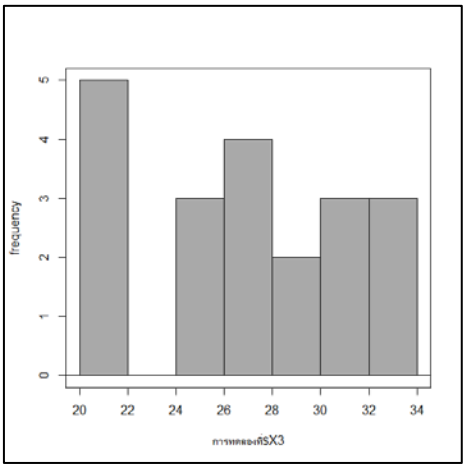
จากแผนภาพการกระจายในตารางที่ 4.6 ช่วยให้เห็นลักษณะความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยนด้วยการสุ่มข้อมูลโดยควบคุมขอบเขตในแต่ละการทดสอบได้ง่ายขึ้น ซึ่งสามารถแสดงให้เห็นว่าข้อมูลที่ได้มีความสัมพันธ์กับข้อมูลต้นแบบปานกลางจนถึงไม่มีความสัมพันธ์กัน



รูปที่ 4.2 การกระจายของข้อมูลของผลการทดสอบที่ 1



รูปที่ 4.3 การกระจายของข้อมูลของผลการทดสอบที่ 2



รูปที่ 4.4 การกระจายของข้อมูลของผลการทดสอบที่ 3

จากฮิสโทแกรมในรูปที่ 4.2 4.3 และ 4.4 แสดงให้เห็นว่าการกระจายของข้อมูลของข้อมูลต้นฉบับและข้อมูลในแต่ละการทดสอบมีความแตกต่างกัน และไม่ได้เป็นการกระจายแบบปกติ

สำหรับข้อมูลอายุและข้อมูลเกรดเฉลี่ยได้ทำการสุ่มค่าจำนวน 3 ครั้งแบ่งเป็น การทดสอบที่ 1 2 และ 3 ตามลำดับ ได้ผลดังตารางที่ 4.7

ตารางที่ 4.7 ตัวอย่างผลที่ได้จากการสุ่มค่าโดยการควบคุมขอบเขตของข้อมูลสำหรับข้อมูลอายุและเกรดเฉลี่ย

ชุดข้อมูล	การทดสอบที่	ค่าความผิดพลาดในการปิดบัง	ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้	ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน
อายุ	1	0	100	-0.0397
	2	0	100	-0.0265
	3	0	100	-0.0407
เกรดเฉลี่ย	1	0	100	-0.0068
	2	0	100	-0.0121
	3	0	100	-0.2304

จากข้อมูลที่ได้ในตารางที่ 4.7 สามารถสรุปจากข้อมูลได้ว่า ข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตสามารถรักษาความเป็นส่วนตัวของข้อมูลต้นฉบับได้เป็นอย่างดี เนื่องจากค่าความผิดพลาดในการปิดบังเป็นค่า 0 และข้อมูลที่ได้อยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้ทุกค่าเนื่องจากร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้เป็น 100 สำหรับความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลที่ได้จากการสุ่มค่าข้อมูลได้ความสัมพันธ์ในระดับต่ำจนถึงไม่มีความสัมพันธ์ ดังนั้นการสุ่มค่าข้อมูลโดยควบคุมขอบเขตสามารถบรรลุวัตถุประสงค์ของงานวิจัยนี้

4.4.2 การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล

จากที่ได้กล่าวไว้ในบทที่ 3 การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูลเป็นการสุ่มข้อมูลที่มีลักษณะคล้ายกับการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตเป็นอย่างเดียว แต่ในวิธีนี้เพิ่มการควบคุมการกระจายของข้อมูลเพื่อให้สามารถสร้างข้อมูลใหม่ที่ตรงตามความต้องการของผู้ปรับเปลี่ยนข้อมูลในกรณีที่ต้องการให้มีการกระจายของข้อมูลระหว่าง

ข้อมูลต้นฉบับและข้อมูลที่ได้หลังการปรับเปลี่ยนมีความคล้ายหรือเหมือนกัน การทดสอบการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูลอาศัยข้อมูลจากข้อมูลตัวอย่าง โดยทำการทดสอบโดยการสุ่มข้อมูลจำนวน 3 การทดสอบ โดยกำหนดให้

การทดสอบที่ 1 กำหนด x = 10%

การทดสอบที่ 2 กำหนด x = 50%

การทดสอบที่ 3 กำหนด x = 90%

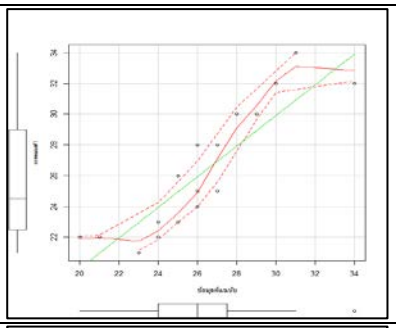
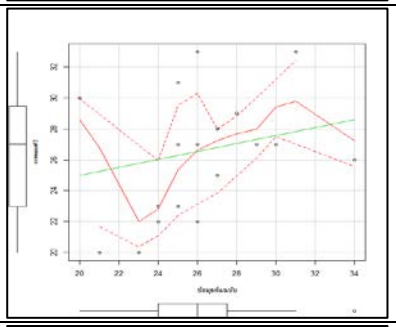
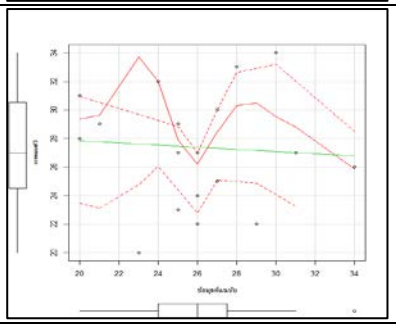
ผลการทดสอบแสดงที่ได้ในแต่ละการทดสอบแสดงในตารางที่ 4.8

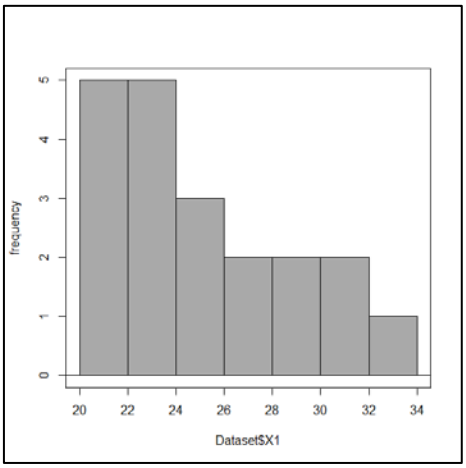
ตารางที่ 4.8 ตัวอย่างของข้อมูลจากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายข้อมูล

แถวที่	**ข้อมูลต้นฉบับ**	**ผลการทดสอบ 1**	**ผลการทดสอบ 2**	**ผลการทดสอบ 3**
1	24	22	22	32
2	29	30	27	22
3	26	25	33	27
4	20	22	30	31
5	26	28	27	27
6	24	23	23	32
7	25	23	27	29
8	26	24	22	22
9	21	22	20	29
10	31	34	33	27
11	30	32	27	34
12	26	24	27	24
13	28	30	29	33
14	23	21	20	20
15	27	28	25	30
16	25	23	31	27
17	34	32	26	26
18	27	25	28	25
19	25	26	23	23
20	20	22	30	28
ค่ามากที่สุด	**34**	**34**	**33**	**34**
ค่าน้อยที่สุด	**20**	**21**	**20**	**20**
ค่าความผิดพลาดในการปิดบัง		0	0	0
ร้อยละของข้อมูลในขอบเขตข้อมูลที่เป็นไปได้		100	100	100
ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน		0.8819	0.2344	-0.0674

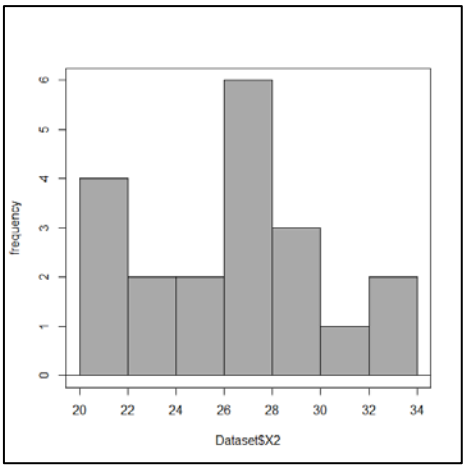
จากตารางที่ 4.8 ข้อมูลที่ผ่านการปรับเปลี่ยนในแต่ละครั้งของการทดสอบสามารถปกปิดข้อมูลต้นฉบับได้เป็นอย่างดีเนื่องจากค่าความผิดพลาดในการปิดบังที่ได้ในแต่ละการทดสอบมีค่าเป็น 0 ทุกครั้ง นอกจากนี้ในการปรับเปลี่ยนข้อมูลด้วยวิธีนี้ ข้อมูลทุกค่าอยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้เนื่องจากมีการควบคุมขอบเขตของข้อมูลไว้เพื่อป้องกันข้อมูลที่สุ่มขึ้นมาอยู่นอกขอบเขตของข้อมูลที่เป็นไปได้ และเมื่อพิจารณาความสัมพันธ์ระหว่างข้อมูลต้นฉบับด้วยค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน จากการทดลองพบว่าความสัมพันธ์ระหว่างข้อมูลขึ้นอยู่กับค่า x ที่ระบุ กล่าวคือ ในการทดลองที่ 1 กำหนดให้ค่า x มีค่า 10% ข้อมูลที่ได้มีความสัมพันธ์สูง เมื่อเพิ่มค่า x ให้เป็นค่า 50% ข้อมูลที่ได้มีความสัมพันธ์ปานกลาง และเมื่อกำหนดค่า x เป็น 90% ข้อมูลที่ได้มีความสัมพันธ์ต่ำ นั่นคือ ยิ่งค่า x มากขึ้น ความสัมพันธ์ของข้อมูลที่ได้ก็จะมากขึ้นตามเป็นผลเนื่องมาจากยิ่งค่า x มีค่าน้อย ช่วงข้อมูลที่เป็นไปได้ที่ถูกสุ่มมีจำนวนน้อยและมีความใกล้เคียงกับข้อมูลต้นฉบับมาก แต่เมื่อค่า x มีค่ามากขึ้น ช่วงข้อมูลที่เป็นไปได้ก็จะมีค่ามากขึ้นและมีความใกล้เคียงกับข้อมูลต้นฉบับน้อยลง สำหรับความสามารถในการปกปิดข้อมูลต้นฉบับสามารถปกปิดได้เป็นอย่างดีเนื่องจากค่าความผิดพลาดในการปิดบังเป็น 0 ดังนั้นการปรับเปลี่ยนข้อมูลด้วยวิธีนี้เหมาะสำหรับกรณีที่ต้องการให้ข้อมูลหลังการปรับเปลี่ยนมีการกระจายของข้อมูลคล้ายกับข้อมูลต้นฉบับ ความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลหลังการปรับเปลี่ยนพิจารณาจากแผนภาพการกระจายจากตารางที่ 4.9

ตารางที่ 4.9 แผนภาพการกระจายของตัวอย่างข้อมูลในตารางที่ 4.8

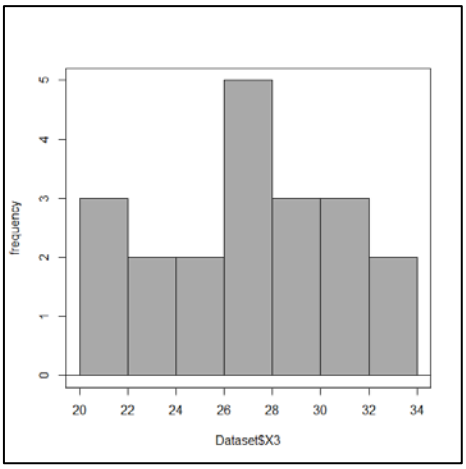
การทดสอบที่	แผนภาพการกระจาย	ความสัมพันธ์ระหว่างชุดข้อมูล
1		ข้อมูลมีความสัมพันธ์กันทางบวก
2		ข้อมูลมีความสัมพันธ์ปานกลาง
3		ข้อมูลมีความสัมพันธ์ต่ำ



รูปที่ 4.5 การกระจายของข้อมูลของผลการทดสอบที่ 1



รูปที่ 4.6 การกระจายของข้อมูลของผลการทดสอบที่ 2



รูปที่ 4.7 การกระจายของข้อมูลของผลการทดสอบที่ 3

จากฮิสโทแกรมในรูปที่ 4.5 4.6 และ 4.7 แสดงให้เห็นว่าลักษณะการกระจายของข้อมูลต้นฉบับและข้อมูลที่ได้จากการปรับเปลี่ยนมีความแตกต่างกัน สำหรับการทดลองที่ 1 และ 2 ในรูปที่ 4.5 และ 4.6 ตามลำดับ เป็นมีการกระจายแบบปกติ แต่การทดลองที่ 3 ในรูปที่ 4.7 ข้อมูลไม่ได้เป็นการกระจายแบบปกติ ดังนั้นลักษณะการกระจายของข้อมูลมีความแตกต่างกันออกไปในแต่ละครั้งของการสุ่มค่าข้อมูล

สำหรับข้อมูลอายุและข้อมูลเกรดเฉลี่ยได้ทำการสุ่มค่าข้อมูลจำนวน 3 การทดสอบ โดยกำหนดให้

- การทดสอบที่ 1 กำหนด x = 10%
- การทดสอบที่ 2 กำหนด x = 50%
- การทดสอบที่ 3 กำหนด x = 90%

ผลการทดสอบแสดงที่ได้ในแต่ละการทดสอบแสดงในตารางที่ 4.10

ตารางที่ 4.10 ตัวอย่างผลที่ได้จากการสุ่มค่าโดยการควบคุมขอบเขตและการกระจายของข้อมูลสำหรับข้อมูลอายุและเกรดเฉลี่ย

ชุดข้อมูล	การทดสอบครั้งที่	ค่าความผิดพลาดในการปิดบัง	ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้	ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน
อายุ	1	0	100	0.8778
	2	0	100	0.1279
	3	0	100	-0.0406
เกรดเฉลี่ย	1	0	100	0.9542
	2	0	100	0.3871
	3	0	100	0.0589

จากข้อมูลที่ได้ในตารางที่ 4.10 สามารถสรุปจากข้อมูลได้ว่า ข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูลสามารถรักษาความเป็นส่วนตัวของข้อมูลต้นฉบับได้เป็นอย่างดีเนื่องจากค่าความผิดพลาดในการปิดบังเป็นค่า 0 และข้อมูลที่ได้อยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้ทุกค่าเนื่องจากร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้เป็น 100 สำหรับความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลที่ได้จากการสุ่มค่าข้อมูลสำหรับข้อมูลอายุได้ความสัมพันธ์ในระดับสูง เมื่อพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันในแต่ละการทดสอบ เช่นเดียวกับการทดลองในตารางที่ 4.8 สามารถสรุปได้ว่าเมื่อค่า x มากขึ้นความสัมพันธ์ของข้อมูลที่ได้ก็จะมากขึ้นตามเป็นผลเนื่องมาจากยิ่งค่า x มีค่าน้อย ช่วงข้อมูลที่เป็นไปได้ที่ถูกสุ่มมีจำนวนน้อยและมีความใกล้เคียงกับข้อมูลต้นฉบับมาก แต่เมื่อค่า x มีค่ามากขึ้น ช่วงข้อมูลที่เป็นไปได้ก็จะมีค่ามากขึ้นและมีความใกล้เคียงกับข้อมูลต้นฉบับน้อยลง ทั้งนี้ขึ้นอยู่กับผู้ทำการปรับเปลี่ยนข้อมูลว่าต้องการให้ข้อมูลมีลักษณะใด ดังนั้นการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูลสามารถบรรลุวัตถุประสงค์ของงานวิจัยนี้และได้ข้อมูลที่มีการกระจายคล้ายกับข้อมูลต้นฉบับ

4.4.3 การสลับที่ของข้อมูล

จากที่ได้กล่าวไว้ในบทที่ 3 การสลับที่ของข้อมูลเป็นการสลับตำแหน่งข้อมูลเพื่อให้ข้อมูลใหม่และข้อมูลเก่ามีการกระจายข้อมูลตรงกัน การทดสอบการสุ่มค่าข้อมูลโดยควบคุมขอบเขตนี้อาศัยข้อมูลจากข้อมูลตัวอย่าง โดยทำการทดสอบโดยการสุ่มข้อมูลจำนวน 3

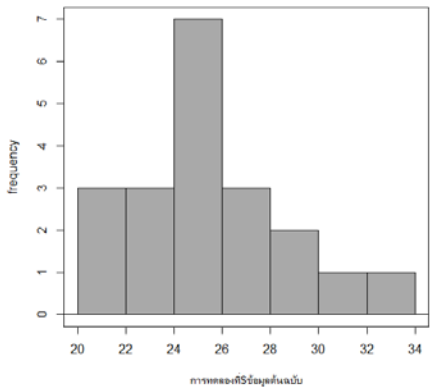
ครั้งแบ่งเป็นการทดสอบที่ 1 2 และ 3 ตามลำดับ ตัวอย่างผลการทดสอบที่ได้แสดงในตารางที่ 4.11

ตารางที่ 4.11 ตัวอย่างของข้อมูลที่ได้จากการสลับที่ของข้อมูล

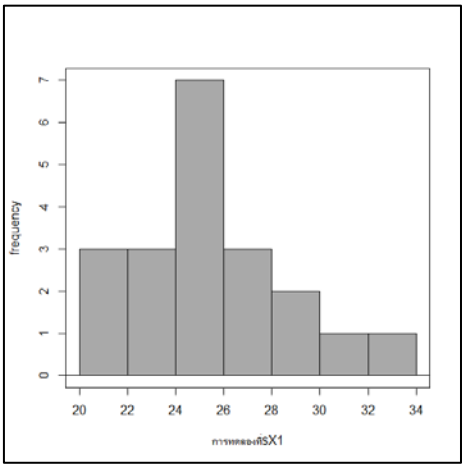
แถวที่	**ข้อมูลต้นฉบับ**	**ผลการทดสอบ 1**	**ผลการทดสอบ 2**	**ผลการทดสอบ 3**
1	24	25	20	25
2	29	31	26	26
3	26	27	24	25
4	20	29	23	27
5	26	24	21	31
6	24	30	34	23
7	25	20	31	24
8	26	25	27	20
9	21	20	26	26
10	31	27	30	30
11	30	26	29	24
12	26	25	27	28
13	28	34	25	21
14	23	24	24	26
15	27	26	28	34
16	25	21	26	26
17	34	26	25	29
18	27	28	20	25
19	25	26	26	20
20	20	23	25	27
ค่ามากที่สุด	**34**	**34**	**34**	**34**
ค่าน้อยที่สุด	**20**	**20**	**20**	**20**
ค่าความผิดพลาดในการปิดบัง		0	0	0
ร้อยละของข้อมูลในขอบเขตข้อมูลที่เป็นไปได้		100	100	100
ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน		0.3321	0.1629	0.1629

จากตารางที่ 4.11 ข้อมูลที่ผ่านการสลับที่ของข้อมูลสามารถปกปิดข้อมูลต้นฉบับได้เป็นอย่างดีเนื่องจากค่าความผิดพลาดในการปิดบังเป็นค่า 0 ทุกครั้งของการสลับที่ของข้อมูล นอกจากนี้ในการปรับเปลี่ยนข้อมูลด้วยวิธีนี้ ข้อมูลทุกค่าอยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้ เนื่องจากมีการข้อมูลทุกค่ายังคงเป็นข้อมูลเดิม และเมื่อพิจารณาความสัมพันธ์ระหว่างข้อมูลต้นฉบับด้วยค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน พบว่า ในแต่ละครั้งของการทดสอบข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยนแล้วมีความสัมพันธ์กันมากน้อยขึ้นอยู่กับการสลับที่ของข้อมูลในแต่ละครั้ง จำนวนข้อมูล และช่วงของข้อมูลที่เป็นไปได้

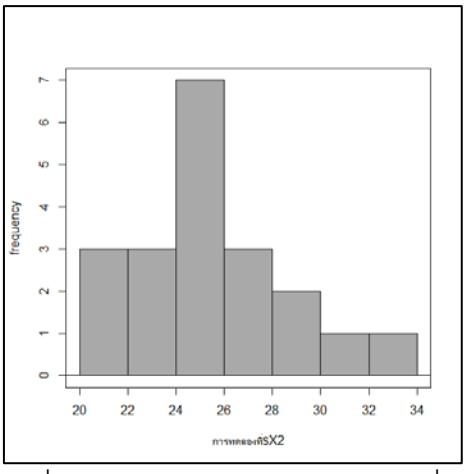
แต่จากที่ได้กล่าวไว้ในบทที่ 3 ว่าวิธีนี้สามารถรักษาการกระจายของข้อมูลไว้ให้คงเดิมเหมือนกับข้อมูลต้นฉบับดังรูปที่ 4.8 สามารถแสดงได้ดังรูปที่ 4.9 4.10 และ 4.11 ตามลำดับ



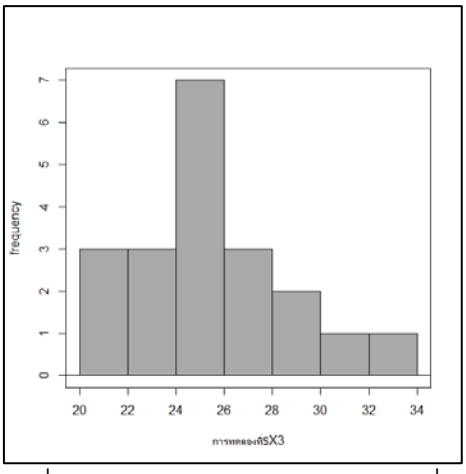
รูปที่ 4.8 ฮิสโทแกรมของข้อมูลต้นฉบับ



รูปที่ 4.9 ฮิสโทแกรมของการทดลองที่ 1



รูปที่ 4.10 ฮิสโทแกรมของการทดลองที่ 2



รูปที่ 4.11 ฮิสโทแกรมของการทดลองที่ 3

4.4.4 การปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุด

จากที่ได้กล่าวไว้ในบทที่ 3 การปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดซึ่งถือว่าเป็นค่าที่ใช้กำหนดขอบเขตที่เป็นไปได้ของข้อมูล การทดสอบการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดอาศัยข้อมูลจากข้อมูลตัวอย่าง โดยทำการทดสอบสุ่มข้อมูลจำนวน 4 การทดสอบ โดยกำหนดให้

- การทดสอบที่ 1 ค่ามากที่สุด = 36 ,ค่าน้อยที่สุด = 18
- การทดสอบที่ 2 ค่ามากที่สุด = 36 ,ค่าน้อยที่สุด = 22
- การทดสอบที่ 3 ค่ามากที่สุด = 32 ,ค่าน้อยที่สุด = 18
- การทดสอบที่ 4 ค่ามากที่สุด = 32 ,ค่าน้อยที่สุด = 22

ทำการแบ่งการทดสอบออกเป็น 2 วิธี คือ การสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล ได้ผลการทดสอบดังตารางที่ 4.11 และ 4.12 ตามลำดับ

ตารางที่ 4.12 การปรับเปลี่ยนค่าขอบเขตร่วมกับการสุ่มค่าข้อมูลโดยการควบคุมขอบเขต

แถวที่	**ข้อมูลต้นฉบับ**	**ผลการทดสอบ 1**	**ผลการทดสอบ 2**	**ผลการทดสอบ 3**	**ผลการทดสอบ 4**
1	24	20	34	32	23
2	29	20	35	21	31
3	26	32	23	22	29
4	20	33	22	27	27
5	26	22	34	24	25
6	24	19	35	22	30
7	25	24	31	32	32
8	26	18	34	20	32
9	21	24	35	23	25
10	31	35	23	25	32
11	30	22	33	27	22
12	26	29	35	25	31
13	28	20	30	22	29
14	23	35	31	28	30
15	27	24	29	18	25
16	25	36	22	29	26
17	34	18	23	32	28
18	27	23	36	22	29
19	25	26	28	24	29
20	20	23	25	25	28
ค่ามากที่สุด	**34**	**36**	**36**	**32**	**32**
ค่าน้อยที่สุด	**20**	**18**	**22**	**18**	**22**
ค่าความผิดพลาดในการปิดบัง		0	0	0	0
ร้อยละข้อมูลในขอบเขตข้อมูลที่เป็นไปได้		100	100	100	100
สัมประสิทธิ์สหสัมพันธ์เพียร์สัน		-0.2368	-0.0532	0.0258	0.1167

ตารางที่ 4.13 การปรับเปลี่ยนค่าขอบเขตร่วมกับการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล กำหนด x=10%

แถวที่	**ข้อมูลต้นฉบับ**	**การทดสอบ 1**	**การทดสอบ 2**	**การทดสอบ 3**	**การทดสอบ 4**
1	24	26	22	23	25
2	29	31	30	31	30
3	26	25	27	28	24
4	20	18	22	18	22
5	26	27	27	28	25
6	24	23	22	25	26
7	25	27	24	26	23
8	26	28	27	25	28
9	21	23	23	19	22
10	31	30	29	28	28
11	30	33	32	28	28
12	26	27	28	25	24
13	28	30	29	29	26
14	23	22	24	22	25
15	27	29	26	29	28
16	25	24	27	26	26
17	34	36	31	32	32
18	27	25	28	28	28
19	25	27	24	24	26
20	20	19	22	19	22
ค่ามากที่สุด	**34**	**33**	**32**	**32**	**32**
ค่าน้อยที่สุด	**20**	**19**	**22**	**18**	**22**
ค่าความผิดพลาดในการปิดบัง		0	0	0	0
ร้อยละข้อมูลในขอบเขตที่เป็นไปได้		100	100	100	100
สัมประสิทธิ์สหสัมพันธ์เพียร์สัน		0.9367	0.8821	0.9076	0.8672

จากผลการทดสอบการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดทั้งในกรณีร่วมกับการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูลดังในตารางที่ 4.11 และ 4.12 ตามลำดับ พบว่า ทุกการทดสอบของกรณีการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและกรณีการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล ได้ผลคือ ค่ามากที่สุดและน้อยที่สุดมีการเปลี่ยนแปลง ค่าความผิดพลาดในการปิดบังเป็นค่า 0 ค่าร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้เป็นค่า 100 และค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันมีความสัมพันธ์ระดับต่ำในกรณีการสุ่มค่าข้อมูลโดยการควบคุมขอบเขต และค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันมีความสัมพันธ์ระดับสูงในกรณีการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายเช่นเดียวกันกับผลของค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันของทั้งสองวิธีในกรณีที่ไม่ได้ทำการเปลี่ยนแปลงค่ามากที่สุดและน้อยที่สุด

สำหรับข้อมูลอายุได้ทำการสุ่มค่าข้อมูล โดยทำการทดสอบสุ่มข้อมูลจำนวน 4 การทดสอบ โดยกำหนดให้

- การทดสอบที่ 1 ค่ามากที่สุด = 40 ,ค่าน้อยที่สุด = 10
- การทดสอบที่ 2 ค่ามากที่สุด = 40 ,ค่าน้อยที่สุด = 25
- การทดสอบที่ 3 ค่ามากที่สุด = 27 ,ค่าน้อยที่สุด = 25
- การทดสอบที่ 4 ค่ามากที่สุด = 27 ,ค่าน้อยที่สุด = 10

ตารางที่ 4.14 ผลการทดสอบการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดของข้อมูลอายุ

วิธีการ	การทดลองที่	ค่ามากที่สุด-ค่าน้อยที่สุดที่ได้	ค่าความผิดพลาดในการปิดบัง	ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้	ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน
การสุ่มค่าข้อมูลโดยควบคุมขอบเขต	1	40-10	0	100	-0.0046
	2	40-25	0	100	0.0200
	3	27-25	0	100	-0.0298
	4	27-10	0	100	0.0029
การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล X = 10%	1	33-17	0	100	0.8281
	2	33-25	0	100	0.7505
	3	27-25	0	100	0.6460
	4	27-17	0	100	0.7795

สำหรับข้อมูลเกรดเฉลี่ยได้ทำการสุ่มค่าข้อมูล โดยทำการทดสอบสุ่มข้อมูลจำนวน 4 การทดสอบ โดยกำหนดให้

- การทดสอบที่ 1 ค่ามากที่สุด = 4.50 ,ค่าน้อยที่สุด = 0.00
- การทดสอบที่ 2 ค่ามากที่สุด = 4.50 ,ค่าน้อยที่สุด = 2.00
- การทดสอบที่ 3 ค่ามากที่สุด = 3.50 ,ค่าน้อยที่สุด = 2.00
- การทดสอบที่ 4 ค่ามากที่สุด = 3.50 ,ค่าน้อยที่สุด = 0.00

ตารางที่ 4.15 ผลการทดสอบการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดของข้อมูลอายุ

วิธีการ	การทดลองที่	ค่ามากที่สุด-ค่าน้อยที่สุดที่ได้	ค่าความผิดพลาดในการปิดบัง	ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้	ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน
การสุ่มค่าข้อมูลโดยควบคุมขอบเขต	1	4.50-0.00	0	100	-0.2482
	2	4.50-2.50	0	100	-0.4613
	3	3.50-2.00	0	100	-0.0180
	4	3.50-0.00	0	100	0.0127
การสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล X = 10%	1	4.50-0.18	0	100	0.6738
	2	4.49-2.00	0	100	0.5669
	3	3.50-2.00	0	100	0.4098
	4	3.50-0.39	0	100	0.5715

จากผลการทดสอบในตารางที่ 4.13 และ 4.14 แสดงให้เห็นว่าข้อมูลที่ได้หลังจากการปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุดยังคงรักษาความเป็นส่วนตัวของข้อมูลต้นฉบับเนื่องจากค่าความผิดพลาดในการปิดบังของทุกการทดสอบเป็นค่า 0 ข้อมูลทุกค่าอยู่ภายในขอบเขตที่ผู้ใช้กำหนดเนื่องจากร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้ของทุกการทดสอบเป็นค่า 100 และค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันของแต่ละการทดสอบได้ความสัมพันธ์เช่นเดียวกันกับผลของค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันของทั้งสองวิธีในกรณีที่ไม่ได้ทำการเปลี่ยนแปลงค่ามากที่สุดและน้อยที่สุด

4.4.5 การปรับเปลี่ยนค่าผลรวม

จากบทที่ 3 การปรับเปลี่ยนค่าผลรวมเป็นการแก้ไขข้อมูลให้มีผลรวมเป็นไปตามที่ผู้ทำการปรับเปลี่ยนข้อมูลระบุ ในการทดสอบนี้ทำการทดสอบเพื่อตรวจสอบว่าผลรวมที่ได้นั้นตรงกับที่ระบุหรือไม่อย่างไร ประกอบด้วย 4 การทดสอบ ดังนี้

- การทดสอบที่ 1 ค่าผลรวม = 600 ทำการรักษาค่าขอบเขต
- การทดสอบที่ 2 ค่าผลรวม = 400 ทำการรักษาค่าขอบเขต

- การทดสอบที่ 3 ค่าผลรวม = 600 ไม่ทำการรักษาค่าขอบเขต
- การทดสอบที่ 4 ค่าผลรวม = 400 ไม่ทำการรักษาค่าขอบเขต

ตารางที่ 4.16 ตัวอย่างของข้อมูลที่ได้จากการปรับเปลี่ยนค่าผลรวม

แถวที่	**ข้อมูลต้นฉบับ**	**ผลการทดสอบ 1**	**ผลการทดสอบ 2**	**ผลการทดสอบ 3**	**ผลการทดสอบ 4**
1	24	29	18	29	19
2	29	34	23	34	24
3	26	31	20	31	21
4	20	20	14	25	15
5	26	31	20	31	21
6	24	29	18	29	19
7	25	30	19	30	20
8	26	31	20	31	21
9	21	26	15	26	16
10	31	36	25	36	26
11	30	35	24	35	25
12	26	31	20	31	21
13	28	33	22	33	23
14	23	28	17	28	18
15	27	32	21	32	22
16	25	30	19	30	20
17	34	39	34	39	29
18	27	32	21	32	22
19	25	30	19	30	20
20	20	20	14	25	15
ค่ามากที่สุด	**34**	**39**	**34**	**39**	**29**
ค่าน้อยที่สุด	**20**	**20**	**14**	**25**	**15**
ค่าผลรวม	**517**	**607**	**403**	**617**	**417**

จากการผลการทดสอบในตารางที่ 4.15 พบว่า

- การทดสอบที่ 1 และ 2 ได้ค่าผลรวมตามที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา (หรืออาจใกล้เคียง ซึ่งเปลี่ยนแปลงตามข้อมูลที่สุ่มได้ในแต่ละครั้ง) และสามารถรักษาค่าน้อยที่สุดค่าเดิมไว้ได้ตามที่ได้กล่าวไว้ในบทที่ 3 ค่าความผิดพลาดในการปิดบังข้อมูลเป็น 0 สำหรับร้อยละของข้อมูลที่อยู่ในขอบเขตได้ร้อยละที่น้อยกว่า 100 เนื่องจากกรณีนี้ต้องมีการบวกเพิ่มค่าผลต่างของค่าเฉลี่ยให้กับข้อมูลแต่ละตัว ทำให้ข้อมูลมีค่ามากกว่าค่าขอบเขตของข้อมูล
- การทดสอบที่ 3 และ 4 ได้ค่าผลรวมตามที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา (หรืออาจใกล้เคียง ซึ่งเปลี่ยนแปลงตามข้อมูลที่สุ่มได้ในแต่ละครั้ง) แต่ค่าขอบเขตได้ถูกเปลี่ยนไปมากน้อยขึ้นอยู่กับค่าผลรวมที่ระบุเข้ามา สำหรับค่าความผิดพลาดในการปิดบังข้อมูลเป็น 0 และร้อยละของข้อมูลที่อยู่ในขอบเขตได้ร้อยละที่น้อยกว่า 100 ด้วยเหตุผลเดียวกับการทดสอบที่ 1 และ 2

สำหรับข้อมูลอายุได้ทำการสุ่มค่าข้อมูลด้วยวิธีการสุ่มค่าโดยควบคุมขอบเขตและการสุ่มค่าโดยควบคุมขอบเขตและการกระจายของข้อมูลร่วมกับการปรับเปลี่ยนค่าผลรวม โดยทำการทดสอบสุ่มข้อมูลจำนวน 4 การทดสอบ โดยกำหนดให้

- การทดสอบที่ 1 ค่าผลรวม = 270000 ทำการรักษาค่าขอบเขต
- การทดสอบที่ 2 ค่าผลรวม = 250000 ทำการรักษาค่าขอบเขต
- การทดสอบที่ 3 ค่าผลรวม = 270000 ไม่ทำการรักษาค่าขอบเขต
- การทดสอบที่ 4 ค่าผลรวม = 250000 ไม่ทำการรักษาค่าขอบเขต

ตารางที่ 4.17 ผลการทดสอบการปรับเปลี่ยนค่าผลรวมร่วมกับการสุ่มค่าทั้งสองแบบสำหรับข้อมูลอายุ

วิธีการ	การทดสอบที่	ค่าผลรวมที่ได้	เก็บรักษาค่าขอบเขตหรือไม่	ค่าความผิดพลาดในการปิดบัง	ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้	ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน
การสุ่มค่าโดยควบคุมขอบเขต	1	268194	เก็บ	0.07	81	-0.0914
	2	240765	เก็บ	0.08	91	-0.1099
	3	279669	ไม่เก็บ	0.08	73	-0.0919
	4	249851	ไม่เก็บ	0	100	-0.0867
การสุ่มค่าโดยควบคุมขอบเขตและการกระจายของข้อมูล X = 10%	1	268965	เก็บ	0.15	85	0.7528
	2	240536	เก็บ	0.16	94	0.7577
	3	270359	ไม่เก็บ	0.16	85	0.7680
	4	240134	ไม่เก็บ	0.19	94	0.7677

สำหรับข้อมูลเกรดเฉลี่ยได้ทำการสุ่มค่าข้อมูลด้วยวิธีการสุ่มค่าโดยควบคุมขอบเขตและการสุ่มค่าโดยควบคุมขอบเขตและการกระจายของข้อมูลร่วมกับการปรับเปลี่ยนค่าผลรวม โดยทำการทดสอบสุ่มข้อมูลจำนวน 4 การทดสอบ โดยกำหนดให้

- การทดสอบที่ 1 ค่าผลรวม = 30000 ทำการรักษาค่าขอบเขต
- การทดสอบที่ 2 ค่าผลรวม = 25000 ทำการรักษาค่าขอบเขต
- การทดสอบที่ 3 ค่าผลรวม = 30000 ไม่ทำการรักษาค่าขอบเขต
- การทดสอบที่ 4 ค่าผลรวม = 25000 ไม่ทำการรักษาค่าขอบเขต

ตารางที่ 4.18 ผลการทดสอบการปรับเปลี่ยนค่าผลรวมร่วมกับการสุ่มค่าทั้งสองแบบสำหรับข้อมูลเกรดเฉลี่ย

วิธีการ	การทดสอบที่	ค่าผลรวมที่ได้	เก็บรักษาค่าขอบเขตหรือไม่	ค่าความผิดพลาดในการปิดบัง	ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้	ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน
การสุ่มค่าโดยควบคุมขอบเขต	1	30000.11	เก็บ	0.0034	75	0.0067
	2	25000.05	เก็บ	0.0025	87	0.0049
	3	30000.14	ไม่เก็บ	0.0028	75	0.0044
	4	25000.28	ไม่เก็บ	0.0027	87	-0.0084
การสุ่มค่าโดยควบคุมขอบเขตและการกระจายของข้อมูล X = 10%	1	29999.91	เก็บ	0.0058	92	0.9315
	2	24999.72	เก็บ	0.0047	100	0.6294
	3	30000.30	ไม่เก็บ	0.0051	92	0.6341
	4	24999.80	ไม่เก็บ	0.0059	99	0.6326

จากผลการทดสอบในตารางที่ 4.17 และ 4.18 แสดงให้เห็นว่าการปรับเปลี่ยนค่าผลรวมสามารถให้ค่าผลรวมได้ตรงกันหรือใกล้เคียงกับค่าที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา สำหรับค่าความผิดพลาดในการปิดบังมีค่าใกล้เคียง 0 ซึ่งถือว่ายังมีความสามารถในการรักษาไว้ซึ่งความเป็นส่วนตัวของข้อมูลต้นฉบับได้ สำหรับร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่ได้เปลี่ยนแปลงทุกครั้งในแต่ละครั้งของการสุ่มค่า แต่โดยรวมแล้วค่าที่ได้อยู่ในช่วง 75-100 นั่นคือข้อมูลส่วนใหญ่อยู่ภายในขอบเขต แต่มีเพียงบางข้อมูลที่อยู่นอกขอบเขตที่เป็นไปได้เนื่องจากข้อมูลจะถูกบวกเพิ่มหรือลบออก ในบางครั้งจำเป็นต้องเกินค่าขอบเขตเพื่อให้ค่าผลรวมได้ตามที่ผู้ทำการปรับเปลี่ยนต้องการ และสำหรับค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันมีลักษณะความสัมพันธ์เช่นเดียวกันกับการสุ่มค่าโดยควบคุมขอบเขตและการสุ่มค่าโดยการควบคุมขอบเขตและการกระจายของข้อมูล

4.4.6 การจัดการค่าว่าง

ในการทดสอบการจัดการค่าว่างนี้ เป็นแค่การแสดงให้เห็นลักษณะของผลที่ได้จากการจัดการค่าว่างในกรณีต่างๆ โดยกำหนด

- การทดสอบที่ 1 รักษาจำนวนของค่าว่างไว้ให้เหมือนต้นฉบับ แต่เปลี่ยนตำแหน่งของค่าว่าง
- การทดสอบที่ 2 กำจัดค่าว่างที่มีอยู่ทั้งหมดโดยการแทนที่ค่าว่างด้วยค่าข้อมูลที่เหมาะสม
- การทดสอบที่ 3 กำหนดค่าว่างเป็นจำนวนตามที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา โดยระบุจำนวนค่าว่าง = 10
- การทดสอบที่ 4 กำหนดค่าว่างโดยสุ่มค่าจำนวนของค่าว่างภายในช่วงที่กำหนด โดยระบุช่วงของจำนวนค่าว่างเป็น 5-10

ตารางที่ 4.19 ตัวอย่างข้อมูลที่ได้จากการจัดการค่าว่าง

แถวที่	ข้อมูลต้นแบบ	ผลการทดสอบ 1	ผลการทดสอบ 2	ผลการทดสอบ 3	ผลการทดสอบ 4
1	21	21	21	21	21
2	42	42	42	42	NULL
3	NULL	48	48	NULL	48
4	48	9	9	9	9
5	9	NULL	82	NULL	82
6	NULL	83	83	NULL	83
7	82	88	88	NULL	88
8	83	NULL	6	NULL	6
9	NULL	1	1	1	1
10	88	79	79	79	NULL
11	6	NULL	85	85	85
12	1	NULL	80	80	NULL
13	NULL	81	81	NULL	NULL
14	79	87	87	NULL	87
15	85	51	51	51	NULL
16	80	82	51	21	80
17	NULL	85	88	NULL	81
18	81	6	1	NULL	48
19	87	80	87	21	80
20	51	NULL	9	NULL	NULL

จากผลการทดสอบในตารางที่ 4.19 จำนวนและตำแหน่งของค่าว่างที่เกิดขึ้นมีความแตกต่างกันออกไปขึ้นอยู่กับการกำหนดของผู้ทำการปรับเปลี่ยน ทั้งนี้อยู่ที่ความต้องการของผู้ทำการปรับเปลี่ยนว่าต้องการให้ข้อมูลที่ได้ออกมาในลักษณะใด

4.4.7 การจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม

การทดสอบการจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยมนี้เป็นการแสดงให้เห็นถึงผลลัพธ์ที่ได้ว่ามีลักษณะอย่างไร โดยกำหนดให้

- การทดสอบที่ 1 กำหนดตัวเลขหลังทศนิยมตามที่ผู้ทำการปรับเปลี่ยนข้อมูลระบุ โดยระบุจำนวนทศนิยม 2 หลัก
- การทดสอบที่ 2 กำหนดตัวเลขหลังทศนิยมโดยการสุ่มค่าตัวเลขในช่วงที่ผู้ทำการปรับเปลี่ยนข้อมูลกำหนดขึ้นมา โดยระบุช่วงของจำนวนทศนิยม 1-4 หลัก

ตารางที่ 4.20 ตัวอย่างข้อมูลที่ได้จากการจัดการตัวเลขหลังจุดทศนิยม

แถวที่	ข้อมูลต้นฉบับ	ผลการทดสอบ 1	ผลการทดสอบ 2
1	9.790	9.79	9.8
2	1.702	1.70	1.7020
3	1.028	1.03	1.0280
4	7.343	7.34	7.3430
5	9.641	9.64	9.641
6	9.644	9.64	9.6440
7	5.824	5.82	5.824
8	8.414	8.41	8.414
9	2.882	2.88	2.882
10	3.925	3.93	3.925
11	2.879	2.88	2.879
12	1.390	1.39	1.390
13	9.549	9.55	9.549
14	0.243	0.24	0.2
15	2.824	2.82	2.8240
16	2.901	2.90	2.90
17	4.725	4.73	4.7
18	0.773	0.77	0.7730
19	1.883	1.88	1.883
20	5.402	5.40	5.4

จากผลการทดสอบในตารางที่ 4.20 ลักษณะของข้อมูลและจำนวนทศนิยมที่ได้ขึ้นอยู่กับการกำหนดของผู้ทำการปรับเปลี่ยน ทั้งนี้อยู่ที่ความต้องการของผู้ทำการปรับเปลี่ยนว่าต้องการให้ข้อมูลที่ได้ออกมาในลักษณะใด

4.4.8 การจัดการข้อมูลที่มีรูปแบบตรงกัน

การทดสอบการจัดการข้อมูลที่มีรูปแบบตรงกันนี้เป็นการแสดงให้เห็นถึงผลลัพธ์ที่ได้ว่ามีลักษณะอย่างไร โดยกำหนดให้

- การทดสอบที่ 1 ทำการสุ่มค่าข้อมูลโดยควบคุมขอบเขต
- การทดสอบที่ 2 ทำการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล

ตารางที่ 4.21 ตัวอย่างข้อมูลที่ได้จากการจัดการข้อมูลเงินเดือน

แถวที่	**ข้อมูลต้นฉบับ**	**ผลการทดสอบ 1**	**ผลการทดสอบ 2**
1	20560	10290	20530
2	30650	10600	30680
3	10440	20180	10420
4	60660	10640	60640
5	10230	20730	10250
6	10560	50770	10530
7	20780	20750	20790
8	30870	50930	30840
9	40870	40150	40880
10	50610	20750	50640
11	50840	50830	50860
12	10890	20050	10900
13	10490	40410	10500
14	20220	10240	20210
15	30230	20330	30260
16	50410	20720	50420
17	50110	30280	50080
18	10780	40240	10790
19	20330	50090	20300
20	50110	40340	50100
ค่ามากที่สุด	**60660**	**50930**	**60640**
ค่าน้อยที่สุด	**10230**	**10240**	**10250**
ค่าความผิดพลาดในการปิดบัง		0	0
ร้อยละของข้อมูลในขอบเขตของข้อมูลที่เป็นไปได้		100%	100%
ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน		-0.03823	0.999999

จากผลการทดสอบในตารางที่ 4.21 พบว่า ลักษณะข้อมูลที่ได้จากทั้ง 2 การทดสอบยังคงมีรูปแบบข้อมูลที่เหมือนกับข้อมูลต้นฉบับ นั่นคือ “X0XX0” สำหรับค่าความผิดพลาดในการปิดบัง ร้อยละของข้อมูลที่อยู่ในขอบเขตของข้อมูลที่เป็นไปได้ และค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันเป็นไปตามลักษณะของการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล

ในบทนี้ได้กล่าวถึงวิธีการทดสอบและผลการทดสอบของวิธีการสุ่มค่าที่ได้นำเสนอในงานวิจัยนี้ ในบทถัดไปเป็นการสรุปและวิเคราะห์ผลการทดสอบที่ได้จากงานวิจัยนี้รวมไปถึงข้อเสนอแนะ

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้นำเสนอวิธีการปรับเปลี่ยนข้อมูลตัวเลขโดยอาศัยข้อมูลต้นฉบับเป็นฐานในการปรับเปลี่ยนข้อมูลเพื่อนำข้อมูลที่ได้ไปใช้เป็นข้อมูลภายในระบบจัดการฐานข้อมูลที่ถูกนำมาใช้เป็นฮันนี่พอตโดยมีจุดประสงค์หลักเพื่อทำให้ผู้โจมตีฐานข้อมูลไม่สามารถตระหนักได้ว่าข้อมูลที่ตนเองกำลังมีปฏิสัมพันธ์อยู่ด้วยนั้นไม่ใช่ข้อมูลจริง ส่งผลให้ข้อมูลต้นฉบับขององค์กรได้รับการปกป้องไว้และสามารถรวบรวมข้อมูลเกี่ยวกับการโจมตีเพื่อนำมาสร้างระบบป้องกันให้กับทรัพยากรต่างๆ ให้กับองค์กรอีกด้วย

จากวัตถุประสงค์ของงานวิจัย ข้อมูลที่ผ่านการปรับเปลี่ยนแล้วต้องมีคุณสมบัติ 3 ข้อ ได้แก่ 1) ข้อมูลต้องมีความแนบเนียน 2) ข้อมูลต้องสามารถปกปิดข้อมูลต้นฉบับได้ และ 3) ข้อมูลต้องมีความแตกต่างเชิงสถิติกับข้อมูลต้นฉบับ โดยงานวิจัยนี้มุ่งเน้นเฉพาะการจัดการข้อมูลตัวเลขเพียงเท่านั้นและเพื่อให้เกิดความหลากหลายของข้อมูลที่สร้างขึ้นมา งานวิจัยนี้จึงได้อาศัยการสุ่มค่าข้อมูลเป็นวิธีการหลักในการสร้างข้อมูลโดยควบคุมข้อมูลให้ที่สุ่มขึ้นมาให้อยู่ภายในค่าขอบเขตของข้อมูลที่นำค่ามากที่สุดและน้อยที่สุดจากข้อมูลต้นฉบับมาเป็นขอบเขต รวมไปถึงมีการควบคุมลักษณะการกระจายของข้อมูลเพื่อให้ข้อมูลที่ได้มีความใกล้เคียงกับข้อมูลต้นฉบับด้วย นอกจากนี้ได้นำเสนอวิธีการปรับเปลี่ยนค่าขอบเขต ค่าผลรวมของชุดข้อมูล การจัดการรูปแบบของข้อมูลด้วยการจัดการจำนวนค่าว่างในชุดข้อมูล การจัดการจำนวนตัวเลขที่แสดงหลังจุดทศนิยม และการจัดการข้อมูลที่มีรูปแบบตรงกัน จากผลการทดสอบในบทที่ 4 พบว่า

- การสุ่มค่าข้อมูลโดยการควบคุมขอบเขตสามารถสร้างข้อมูลที่ตรงตามวัตถุประสงค์ทั้ง 3 ข้อ โดยที่ข้อมูลมีความแนบเนียนเพราะข้อมูลทุกตัวอยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้ ข้อมูลต้นฉบับมีความปลอดภัยเพราะมีค่าความผิดพลาดเป็น 0 และเมื่อพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์

ของเพียร์สัน ความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลหลังการปรับเปลี่ยนมีความสัมพันธ์กันต่ำถึงไม่มีความสัมพันธ์เลย

- การสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายข้อมูลให้ข้อมูลที่มีความแนบเนียนและปลอดภัยเช่นเดียวกับการสุ่มค่าข้อมูลโดยการควบคุมขอบเขต แต่ความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลที่ได้มีความสัมพันธ์กันค่อนข้างสูง เนื่องจากวิธีนี้ต้องการให้การกระจายของข้อมูลมีค่าใกล้เคียงกัน ความสัมพันธ์ของข้อมูลทั้งสองจึงมีความสัมพันธ์กันด้วย
- การสลับที่ของข้อมูลให้ข้อมูลที่มีความแนบเนียนและปลอดภัย แต่ความสัมพันธ์ระหว่างข้อมูลมีการเปลี่ยนแปลงไปตามลักษณะของข้อมูลและการสลับที่ในแต่ละครั้ง แต่สิ่งที่ได้คือการกระจายของข้อมูลมีลักษณะเหมือนกับข้อมูลต้นฉบับ
- การปรับเปลี่ยนค่ามากที่สุดและน้อยที่สุด พบว่า ข้อมูลที่ได้มีการเปลี่ยนแปลงของค่ามากที่สุดและน้อยที่สุด ค่าความผิดพลาดในการปิดบังเป็น 0 ข้อมูลมีความแนบเนียนเพราะข้อมูลทุกตัวอยู่ภายในขอบเขตของข้อมูลที่เป็นไปได้ และค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันและลักษณะของกราฟเส้นเป็นไปตามลักษณะของข้อมูลที่ได้จากการสุ่มค่าข้อมูลแต่ละวิธี
- การปรับเปลี่ยนค่าผลรวมสามารถปกปิดข้อมูลต้นฉบับได้เช่นกัน แต่อาจส่งผลให้ข้อมูลมีการออกนอกขอบเขตของข้อมูลที่เป็นไปได้มากหรือน้อยแตกต่างกันไปในแต่ละครั้งที่ทำการสุ่มค่าข้อมูล แต่สำหรับความสัมพันธ์ระหว่างข้อมูลต้นฉบับและข้อมูลที่ได้หลังการปรับเปลี่ยนพบว่ากรณีรักษาค่าขอบเขตเดิมไว้ข้อมูลทั้งสองมีความสัมพันธ์ต่ำ แต่กรณีไม่รักษาค่าขอบเขตเดิมไว้ข้อมูลทั้งสองมีความสัมพันธ์กันเชิงบวก เนื่องจากถ้ารักษาค่าขอบเขตเดิมไว้ยังคงมีข้อมูลบางค่าที่อยู่ในขอบเขตของข้อมูลต้นฉบับ แต่ถ้าไม่รักษาค่าขอบเขตเดิมไว้เมื่อพิจารณาจากกราฟเส้นพบว่าข้อมูลทุกค่าจะมีลักษณะ

เยื้องขึ้นหรือเยื้องลงในกรณีค่าผลรวมค่าใหม่มากกว่าค่าเดิมและค่าผลรวมค่าใหม่น้อยกว่าค่าเดิมตามลำดับ

- การจัดการจำนวนค่าว่าง จำนวนตัวเลขที่แสดงหลังจุดทศนิยม และข้อมูลที่มีรูปแบบเหมือนกัน สามารถกำหนดและปรับเปลี่ยนชุดข้อมูลให้เป็นไปตามที่ผู้ทำการปรับเปลี่ยนต้องการ

ทั้งนี้ลักษณะข้อมูลที่สร้างขึ้นมาในแต่ละครั้ง ผู้ทำการปรับเปลี่ยนข้อมูลถือเป็นผู้ตัดสินว่าอยากให้ข้อมูลที่ได้ออกมาในลักษณะใดเพราะแต่ละวิธีที่นำเสนอในงานวิจัยนี้ต่างให้ข้อมูลที่แตกต่างกันไป ดังนั้นก่อนที่จะทำการปรับเปลี่ยนข้อมูลควรมีการวางแผนและออกแบบลักษณะข้อมูลที่เหมาะสมกับองค์กร เพราะแต่ละองค์กรล้วนมีความต้องการที่แตกต่างกันออกไป รวมไปถึงข้อมูลที่ได้ในแต่ละวิธีอาจจะมีความเหมาะสมกับองค์กรหนึ่งแต่ไม่เหมาะสมกับอีกองค์กรหนึ่งขึ้นอยู่กับการตัดสินใจของแต่ละองค์กร

5.2 **ข้อเสนอแนะ**

- ในงานวิจัยนี้เป็นเพียงการปรับเปลี่ยนข้อมูลเพื่อนำไปใช้ในฐานข้อมูลที่จะนำไปใช้เป็นฮันนี่พอต มีการทดสอบเพียงคุณภาพของข้อมูลที่ได้เพียงเท่านั้น ขาดการทดสอบในเรื่องของการนำไปใช้งานจริง ในสถานการณ์จริง ดังนั้งในส่วนนี้จึงสามารถทดสอบได้เพื่อช่วยให้ทราบได้ว่าข้อมูลดังกล่าวมีประสิทธิภาพในการดึงดูดผู้โจมตีมากน้อยเพียงใด
- นอกจากวิธีการที่ได้นำเสนอในงานวิทยานิพนธ์เล่มนี้ การใช้วิธีการตรวจจับข้อมูลที่ผ่านการปลอมแปลง แล้วนำข้อบกพร่องของข้อมูลที่ตรวจพบเจอหรือตัวชี้วัดที่บ่งบอกว่าข้อมูลนั้นคือข้อมูลปลอมมาช่วยพัฒนาวิธีการปรับเปลี่ยนข้อมูลให้มีความแนบเนียนและปลอดภัยมากยิ่งขึ้น
- นอกจากนี้ข้อมูลประเภทตัวอักษรที่งานวิจัยนี้ไม่ได้นำเสนอนั้น ผู้ที่สนใจสามารถที่จะพัฒนาโดยการใช้หลักการของการประมวลผลภาษาธรรมชาติ (Natural Language Processing) [29]

- ในการสลับที่ของข้อมูล การกระจายของข้อมูลที่ได้นั้นมีลักษณะเหมือนกับต้นฉบับทั้งรูปร่างและค่าช่วงของข้อมูลในแต่ละแท่งของกราฟ การเลื่อนกราฟจึงเป็นอีกวิธีหนึ่งที่ช่วยเพิ่มความปลอดภัยให้กับข้อมูลต้นฉบับและยังคงไว้ซึ่งลักษณะการกระจายของข้อมูลที่ในวิทยานิพนธ์เล่มนี้ไม่ได้นำเสนอ
- ในเรื่องของการจัดการข้อมูลที่มีรูปแบบ ในวิทยานิพนธ์เล่มนี้สนใจเพียงแต่เลข 0 และไม่ใช่เลข 0 เท่านั้น ผู้สนใจสามารถปรับปรุงโดยการจัดการรูปแบบกรณีที่เป็นเลขอื่นๆ ได้ เช่น กรณีที่เป็น “XX9” นั่นคือ ทุกค่าในชุดข้อมูลมีหลักหน่วยเป็นเลข 9 เป็นต้น

รายการอ้างอิง

[1] Joho, D. Active Honeypots [Online]. 2004. Available from: http://www.ifi.uzh.ch/archive/mastertheses/DA_Arbeiten_2004/Joho_Dieter.pdf [2009, November 19].

[2] Mokube, I. and Adams, M., Honeypots: concepts, approaches, and challenges. Proceedings of the 45th annual southeast regional conference (ACM-SE'45), 2007.

[3] Spitzner, L. Honeypots: Definitions and Value of Honeypots [Online]. 2001. Available from: http://www.securityfocus.com/infocus/1492 [2009, November 13].

[4] Spitzner, L. Honeypots: Tracking Hackers. USA: Addison-Wesley, 2003.

[5] Provos, N. Developments of the Honeyd Virtual Honeypot [Online]. Available from: http://www.honeyd.org [2009, November 20].

[6] NETSEC. Specter: Intrusion Detection System [Online]. Available from: http://www.specter.com/ [2009, November 25].

[7] Honeynet Project & Research Alliance. Know Your Enemy: Honeywall CDROM Roo [Online]. 2005. Available from: http://old.honeynet.org/papers/cdrom/roo/index.html [2009, November 8].

[8] The Honeynet Project. Honeywall CDROM [Online]. Available from: https://projects.honeynet.org/honeywall/ [2009, November 6].

[9] mwcollect Alliance. Collaborative Malware Collection and Sensing [Online]. Available from: http://code.mwcollect.org/ [2009, November 26].

[10] The Honeynet Project. Know Your Enemy: Defining Virtual Honeynets [Online]. 2003. Available from: http://old.honeynet.org/papers/virtual/ [2009, November 18].

[11] The Honeynet Project. Know Your Enemy: Honeynets [Online]. 2006. Available from: http://old.honeynet.org/papers/honeynet/ [2009, November 12].

[12] The Honeynet Project. Know Your Enemy: Learning about Security Threats. 2nd ed. USA: Addison-Wesley, 2004.

[13] Chickowski, E. Why Your Database Are Vulnerable To Attack – And What You Can Do About It [Online]. Available From: http://www.darkreading.com/tech-center/2/Database_Security.html [2011, August 5].

[14] Oltsik, J. Database at Risk. [Online] 2009. Available from: www.enterprisestrategygroup.com/2009/09/esg-research-brief-databases-at-risk/ [2011, August 5].

[15] Rong, C. and Yang, G. Honeypots in Blackhat Mode and its Implications. Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2003), 2003.

[16] Abie, H. An Overview of Firewall Technologies [Online]. 2000. Available from: http://www.nr.no/publications/FirewallTechnologies.pdf [2009, November 11].

[17] Cheswick, W., R., Bellovin, S., M. and Rubin, A., D. Firewalls and Internet Security, Repelling the Wily Hacker. 2nd ed. USA: Addison-Wesley, 2003.

[18] Axelsson, S. Research in intrusion-detection systems: A survey [Online]. 1999. Available from: http://www.cs.uiuc.edu/class/fa05/cs591han/papers/axelssonSurvey99.pdf [2011, August 19].

[19] Agrawal, R., and Srikant, R. Privacy-Preserving Data Mining. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000.

[20] Denning, D.E., Cryptography and Data Security. Newyork: Addison-Wesley, 1982.

[21] Lin, J., Liu and J. Privacy Preserving Itemset Mining Through Fake Transactions. Proceedings of the 2007 ACM symposium on Applied computing, 2007.

[22] Muralidhar, K., and Sarathy, R. Security of Random Data Perturbation Methods. ACM Trans. Database Syst, 1999.

[23] Yao, Y., Huang, L., Yang, W., Luo, Y., Jing, W., and Xu, W. Privacy-preserving Technology and Its Applications in Statistics Measurements. Proceedings of the 2nd international conference on Scalable information systems, 2007.

[24] Yuill, J., Zappe, M., Denning, D., and Feer, F. Honeyfiles: Deceptive Files for Intrusion Detection. Proceedings from the Fifth Annual IEEE SMC, 2004.

[25] Rowe, N., C. Measuring the Effectiveness of Honeypot Counter-Counterdeception. Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006.

[26] Gupta, S., K., Damor, R., Gupta, A., and Goyal, V. OCHD: Preserving Obliviousness Characteristic of Honeypot Database. Proceedings of 13th International Conference on Management of Data, 2006.

[27] Bertino, E., Lin, D., and Jiang, W. A Survey of Quantification of Privacy Preserving Data Mining Algorithms. [Online] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.4396&rep=rep1&type=pdf [2011, August 5].

[28] Oliveira, S.R.M., Zaiane, O.R.. Privacy preserving frequent itemset mining. Proceedings of IEEE icdm Workshop on Privacy, Security and Data Mining, 2002.

[29] Wikipedia. Natural language processing [Online]. Available from: http://en.wikipedia.org/wiki/Natural_language_processing [2012, April 24].

ภาคผนวก

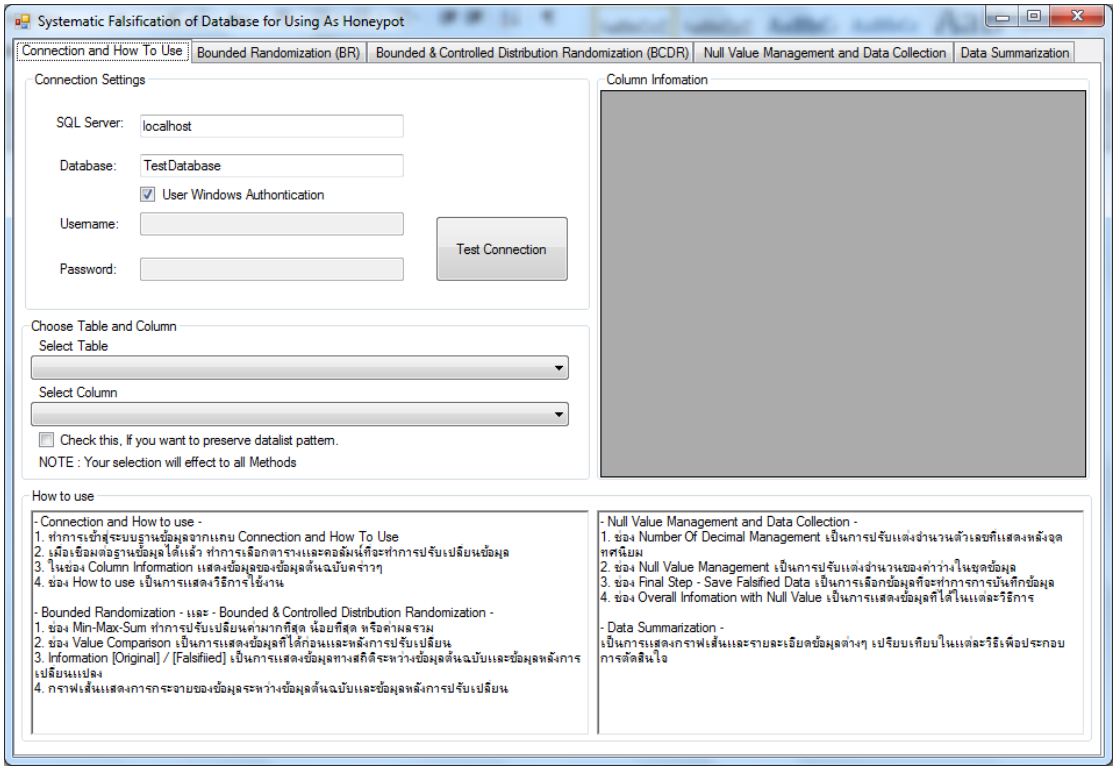
ภาคผนวก ก

โปรแกรมที่ใช้ในการปรับเปลี่ยนข้อมูลที่นำเสนอในงานวิจัย

ในงานวิจัยนี้ ผู้วิจัยได้ทำการเขียนโปรแกรมเพื่ออำนวยความสะดวกในการปรับเปลี่ยนข้อมูลตามวิธีที่ได้นำในบทที่ 3 ในส่วนของภาคผนวก ก อธิบายถึงลักษณะหน้าตาของโปรแกรมและการใช้งานในแต่ละส่วน ประกอบด้วย

1. ลักษณะโปรแกรมและการใช้งาน

1.1 หน้าต่างการเชื่อมต่อฐานข้อมูล รายละเอียดเกี่ยวกับข้อมูลต้นฉบับ และวิธีการใช้งานโปรแกรม Systematic Falsification of Database for Using As Honeypot



รูปที่ ก.1 รายละเอียดของแถบ Connection and How To Use

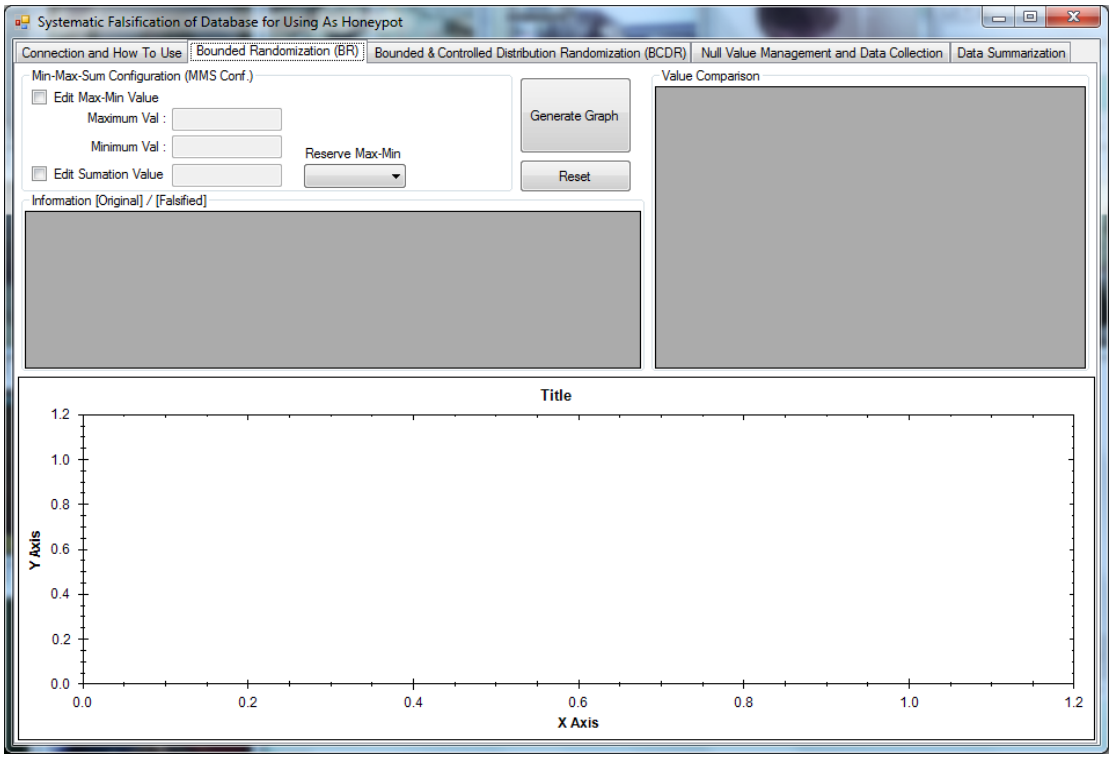
ในหน้าต่างนี้ประกอบด้วย

- Connection Settings เป็นส่วนที่ใช้ในการติดต่อกับฐานข้อมูล โดยที่ผู้ใช้งานต้องทำการกรอกรายละเอียดที่โปรแกรมต้องการ จากนั้นกดปุ่ม Test Connection เพื่อทำการทดสอบว่าสามารถเชื่อมต่อกับฐานข้อมูลได้หรือไม่

ถ้าสามารถเชื่อมต่อได้จะมีหน้าต่างขึ้นมาพร้อมข้อความ “Connection Succeeded” แต่ถ้าไม่สามารถเชื่อมต่อได้จะมีข้อความความผิดพลาดแจ้งให้ผู้ใช้ทราบ

- Choose Table and Column เป็นส่วนที่ใช้สำหรับเลือกตารางและคอลัมน์ที่ต้องการทำการปรับเปลี่ยนข้อมูล รายการตารางและคอลัมน์สามารถเลือกได้เมื่อทำการเชื่อมต่อกับฐานข้อมูลเรียบร้อยแล้ว
- Column Information เป็นส่วนที่แสดงรายละเอียดเกี่ยวกับข้อมูลในตารางและคอลัมน์ที่ถูกเลือก ซึ่งรายละเอียดประกอบด้วยจำนวนแถว ค่ามากที่สุด ค่าน้อยที่สุด ค่าผลรวม ข้อมูลทางสถิติ เป็นต้น
- How to use เป็นส่วนอธิบายการใช้งานว่าแต่ละส่วนทำหน้าที่อะไรได้บ้าง

1.2 หน้าต่างการสุ่มค่าข้อมูลโดยการควบคุมขอบเขต



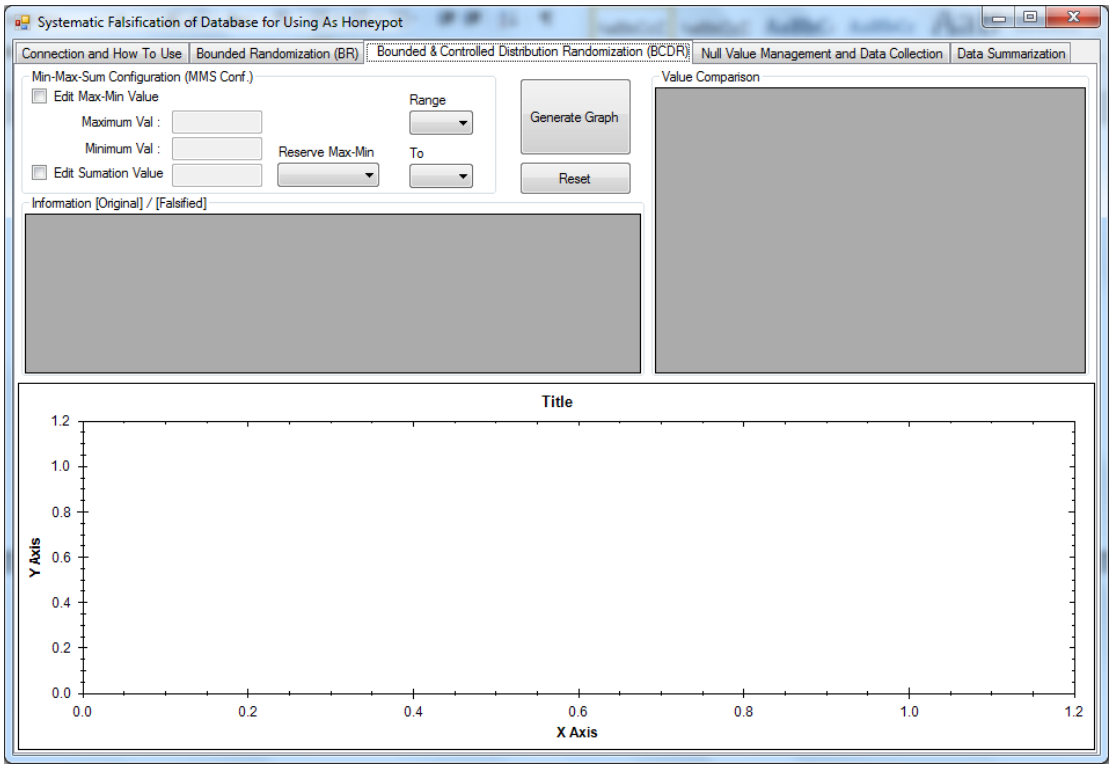
รูปที่ ก.2 รายละเอียดของแถบ Bounded Randomization (BR)

ในหน้าต่างนี้ประกอบด้วย

- Min-Max-Sum Configuration (MMS Conf.) เป็นส่วนที่ใช้สำหรับปรับแต่งค่ามากที่สุด ค่าน้อยที่สุด และค่าผลรวมตามที่ผู้ทำการปรับแต่งต้องการ

- ปุ่ม Generate Graph เป็นปุ่มที่ใช้ในการสุ่มค่าข้อมูลโดยควบคุมขอบเขต เมื่อกดปุ่มแล้ว ระบบจะทำการสุ่มค่าข้อมูลและสร้างกราฟให้เห็นในส่วนของการแสดงกราฟ (พื้นที่สีขาว) โดยกราฟที่ได้เป็นกราฟเส้นประกอบไปด้วยกราฟเส้นของข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยน
- เป็น Reset เป็นปุ่มที่ใช้ล้างข้อมูลที่ได้จากการสุ่มครั้งก่อน
- Value Comparison เป็นส่วนที่แสดงข้อมูลที่ได้โดยเปรียบเทียบกันระหว่างข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยน

1.3 หน้าต่างการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล



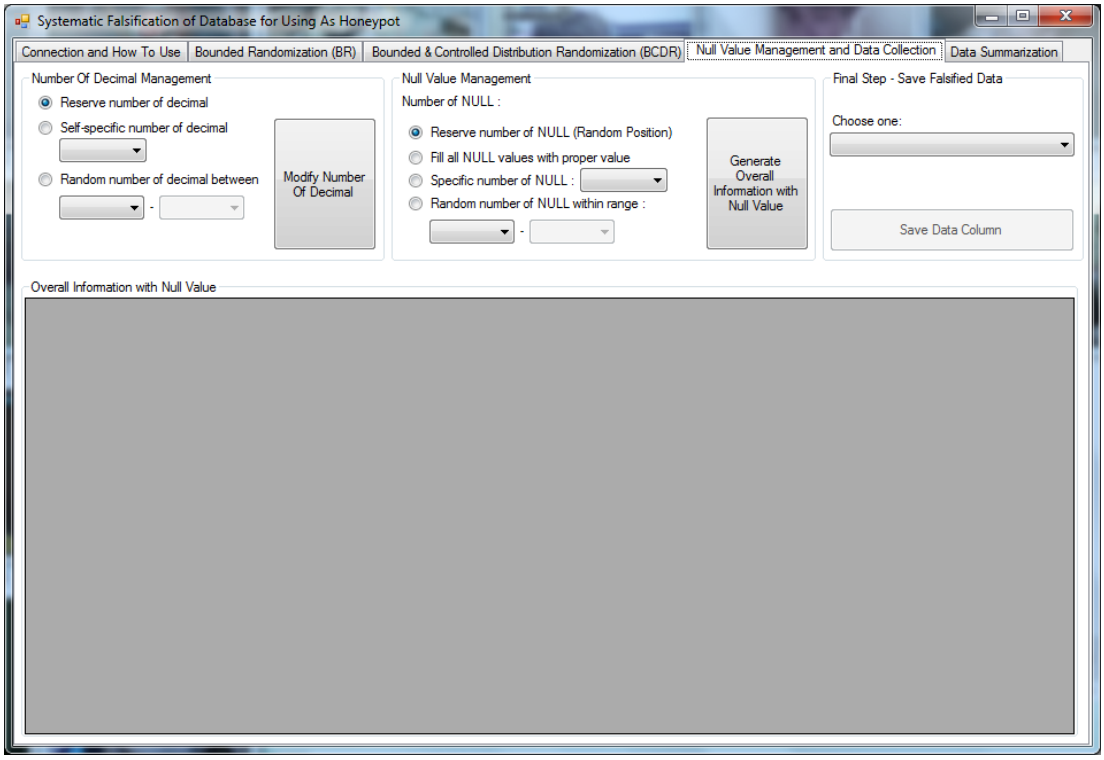
รูปที่ ก.3 รายละเอียดของแถบ Bounded & Controlled Distribution Randomization (BCDR)

ในหน้าต่างนี้ประกอบด้วย

- Min-Max-Sum Configuration (MMS Conf.) เป็นส่วนที่ใช้สำหรับปรับแต่งค่ามากที่สุด ค่าน้อยที่สุด และค่าผลรวมตามที่ผู้ทำการปรับแต่งต้องการ
- Range … to … เป็นส่วนที่ใช้ในการกำหนดช่วงค่า [x, y] ตามวิธีที่ได้นำเสนอในบทที่ 3 โดยที่ค่า x กำหนดให้มีค่า {-1, -2, -3, -4, -5} และ ค่า y กำหนดให้มีค่า {1, 2, 3, 4, 5}

- ปุ่ม Generate Graph เป็นปุ่มที่ใช้ในการสุ่มค่าข้อมูลโดยควบคุมขอบเขต เมื่อกดปุ่มแล้ว ระบบจะทำการสุ่มค่าข้อมูลและสร้างกราฟให้เห็นในส่วนของการแสดงกราฟ (พื้นที่สีขาว) โดยกราฟที่ได้เป็นกราฟเส้นประกอบไปด้วยกราฟเส้นของข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยน
- เป็น Reset เป็นปุ่มที่ใช้ล้างข้อมูลที่ได้จากการสุ่มครั้งก่อน
- Value Comparison เป็นส่วนที่แสดงข้อมูลที่ได้โดยเปรียบเทียบกันระหว่างข้อมูลต้นฉบับและข้อมูลที่ผ่านการปรับเปลี่ยน

1.4 หน้าต่างการปรับแต่งค่าจำนวนของค่าว่าง จำนวนตัวเลขที่แสดงหลังจุดทศนิยม และการบันทึกข้อมูล



รูปที่ ก.4 รายละเอียดแถบ Null Value Management and Data Collection

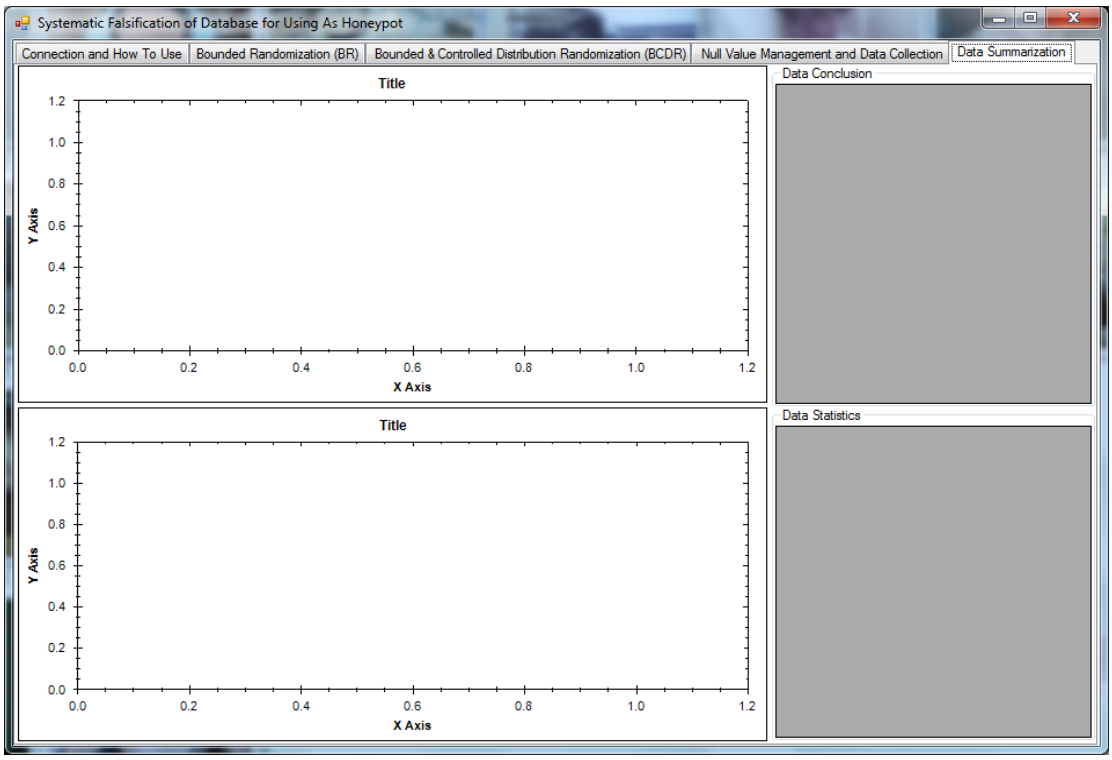
ในหน้าต่างนี้ประกอบด้วย

- Number Of Decimal Management เป็นส่วนของการปรับแต่งจำนวนตัวเลขที่แสดงหลังจุดทศนิยมตามที่ผู้ทำการปรับเปลี่ยนข้อมูลต้องการ โดยมีตัวเลือกดังนี้
  - Reserve Number of decimal ทำการปรับจำนวนตัวเลขที่แสดงหลังจุดทศนิยมให้เป็นไปตามข้อมูลต้นฉบับ

- Self-specific number of decimal ทำการปรับจำนวนตัวเลขที่แสดงหลังจุดทศนิยมให้เป็นไปตามที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา โดยส่งผลต่อข้อมูลทุกค่า
- Random number of decimal between ทำการปรับจำนวนตัวเลขที่แสดงหลังจุดทศนิยมให้เป็นไปตามช่วงที่ผู้ทำการปรับเปลี่ยนระบุเข้ามา โดยส่งผลต่อข้อมูลทุกค่า

- Null Value Management เป็นส่วนของการปรับแต่งจำนวนค่าว่างตามที่ผู้ทำการปรับเปลี่ยนข้อมูลต้องการ โดยมีตัวเลือกดังนี้
  - Reserve number of NULL (Random Position) ทำการกำหนดให้จำนวนของค่าว่างมีค่าเหมือนกับข้อมูลต้นฉบับ แต่ทำการสลับตำแหน่งของค่าว่าง
  - Fill all NULL values with proper value ทำการกำจัดค่าว่างโดยการเติมค่าที่เป็นไปได้เข้าไป
  - Specific number of NULL ทำการกำหนดจำนวนของค่าว่างตามที่ผู้ทำการปรับเปลี่ยนกำหนด
  - Random number of NULL within range ทำการกำหนดช่วงจำนวนของค่าว่างตามที่ผู้ทำการปรับเปลี่ยนกำหนด โดยจำนวนของค่าว่างระบบจะทำการสุ่มภายในช่วงดังกล่าว
- Final Step – Save Falsified Data เป็นส่วนที่ใช้เลือกข้อมูลที่ผู้ทำการปรับเปลี่ยนต้องการเพื่อทำการบันทึกข้อมูลดังกล่าวกลับไปยังฐานข้อมูล
- Overall Information with Null Value เป็นส่วนที่แสดงข้อมูลเปรียบเทียบของแต่ละวิธีการเพื่อให้เห็นถึงข้อมูลที่เกิดขึ้นจริง ใช้ในการประกอบการตัดสินใจ

1.5 แสดงข้อมูลเปรียบเทียบในแต่ละวิธีของการสุ่มค่าข้อมูล



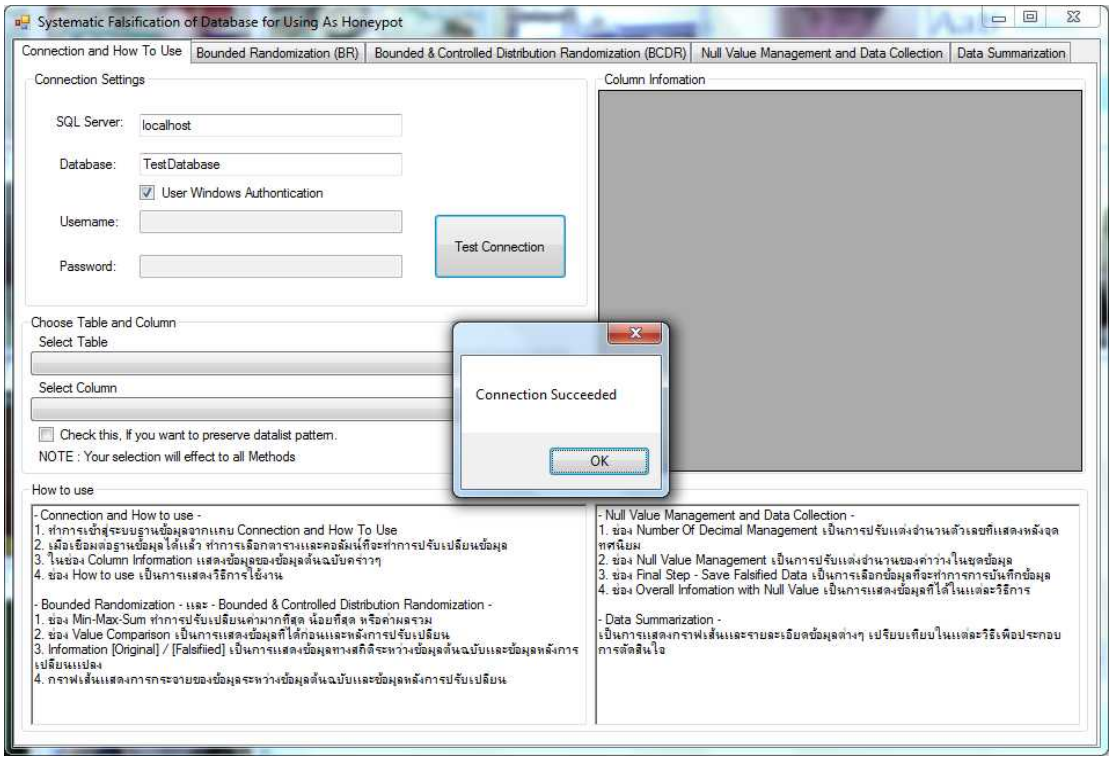
รูปที่ ก.5 รายละเอียดแถบ Data Summarization

ในหน้าต่างนี้ประกอบด้วย

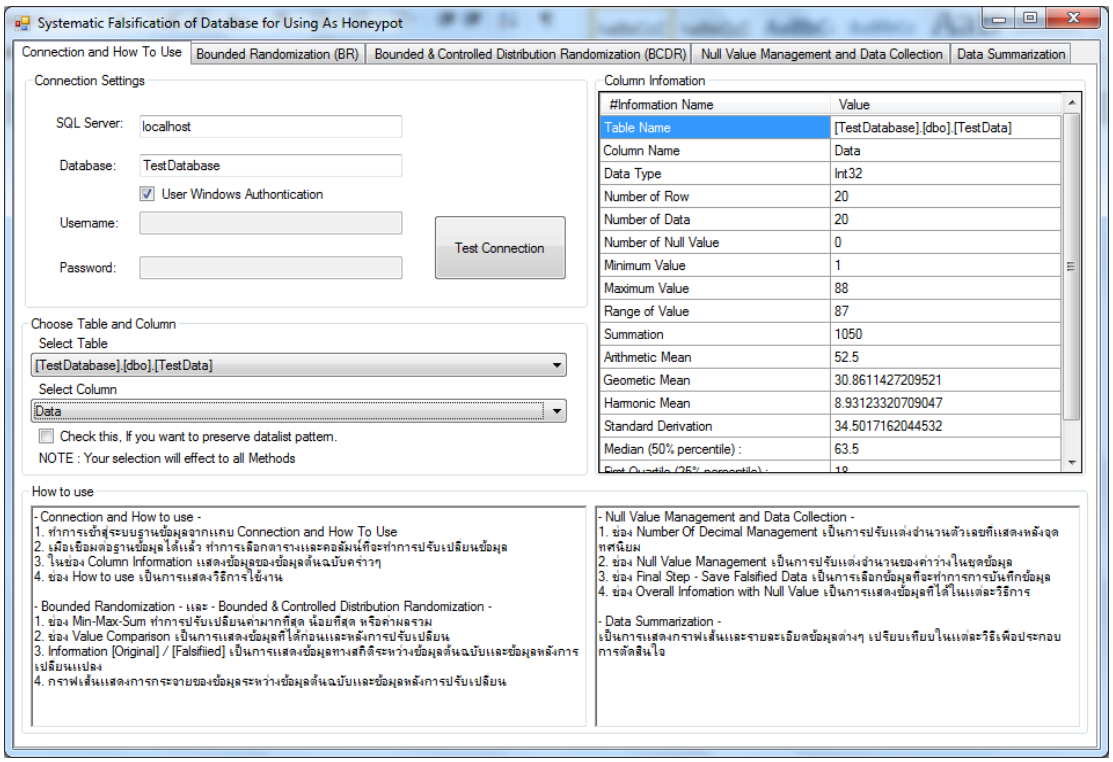
- ส่วนของการแสดงกราฟเส้นโดยที่กรอบด้านบนแสดงกราฟเส้นของการสุ่มค่าข้อมูลโดยควบคุมขอบเขต และกรอบด้านล่างแสดงกราฟเส้นของการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล
- Data Conclusion เป็นส่วนที่ใช้แสดงข้อมูลต้นฉบับ ข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล
- Data Statistic เป็นส่วนที่ใช้แสดงข้อมูลทางสถิติของข้อมูลต้นฉบับ ข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจายของข้อมูล

2. ตัวอย่างการใช้งานโปรแกรม

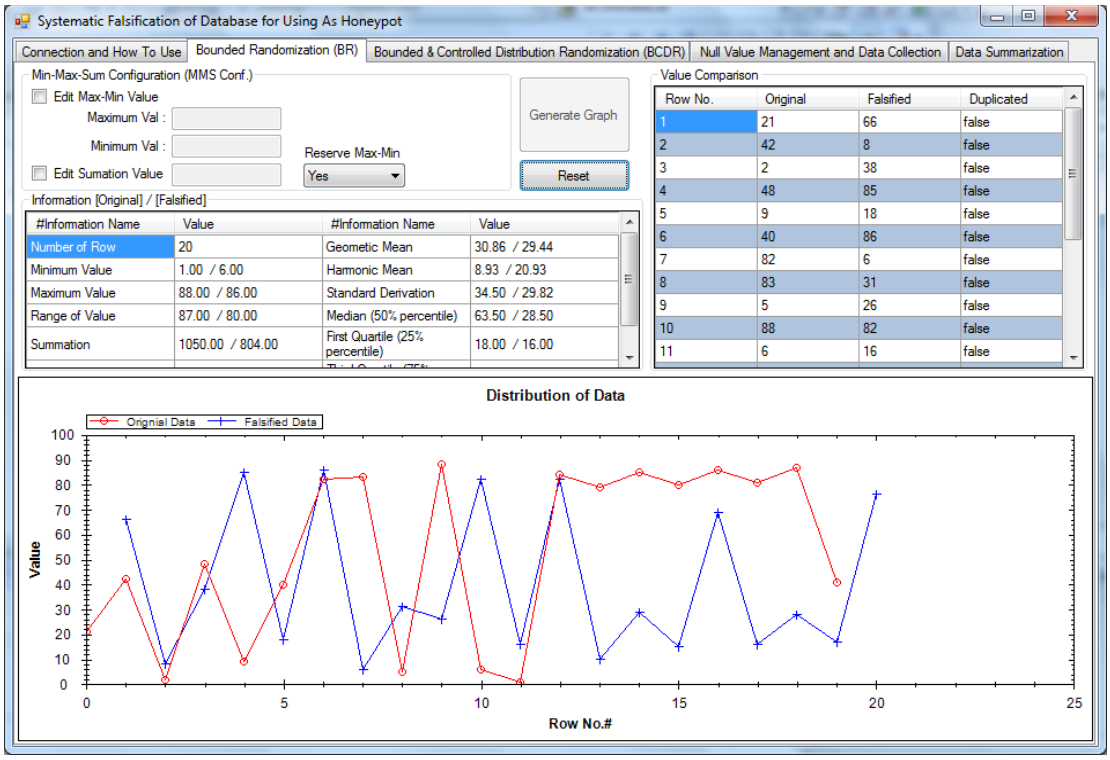
ในส่วนนี้เป็นการแสดงตัวอย่างของการสุ่มค่าข้อมูลในวิธีต่างๆ พร้อมทั้งตัวอย่างข้อมูล ดังแสดงในรูปที่ ก.6 ก.7 ก.8 ก.9 ก.10 และ ก.11 ตามลำดับ



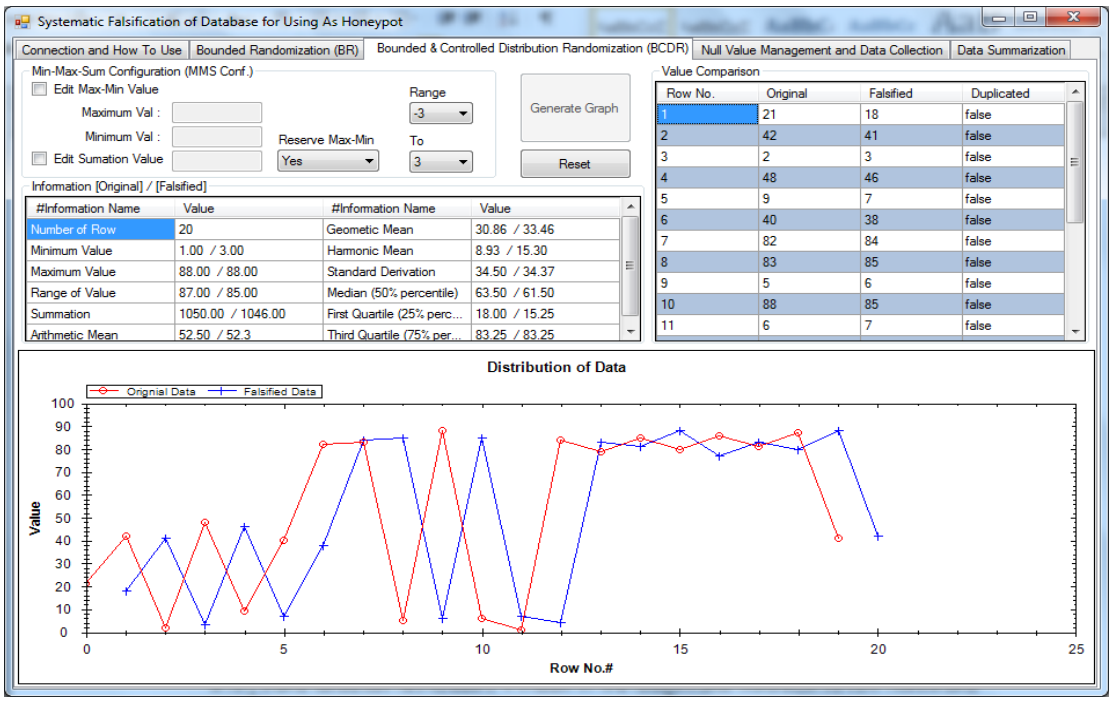
รูปที่ ก.6 ตัวอย่างการเชื่อมต่อฐานข้อมูลสำเร็จ



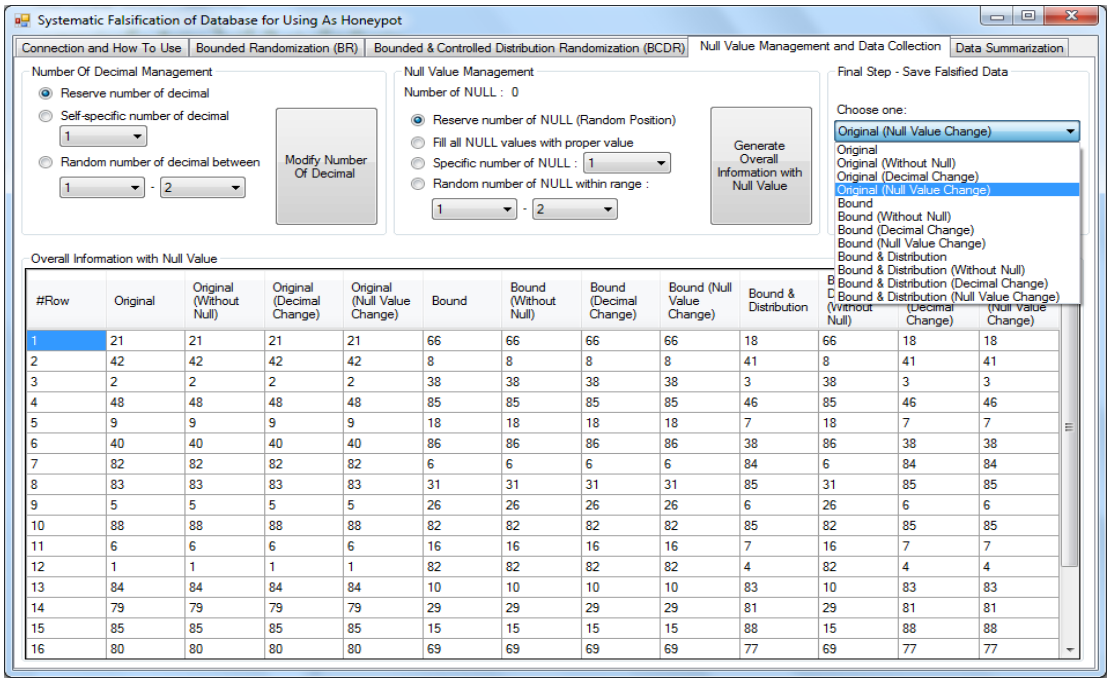
รูปที่ ก.7 แสดงตัวอย่างการเลือกตารางและคอลัมน์ พร้อมด้วยข้อมูลเกี่ยวกับข้อมูลในคอลัมน์ที่เลือก



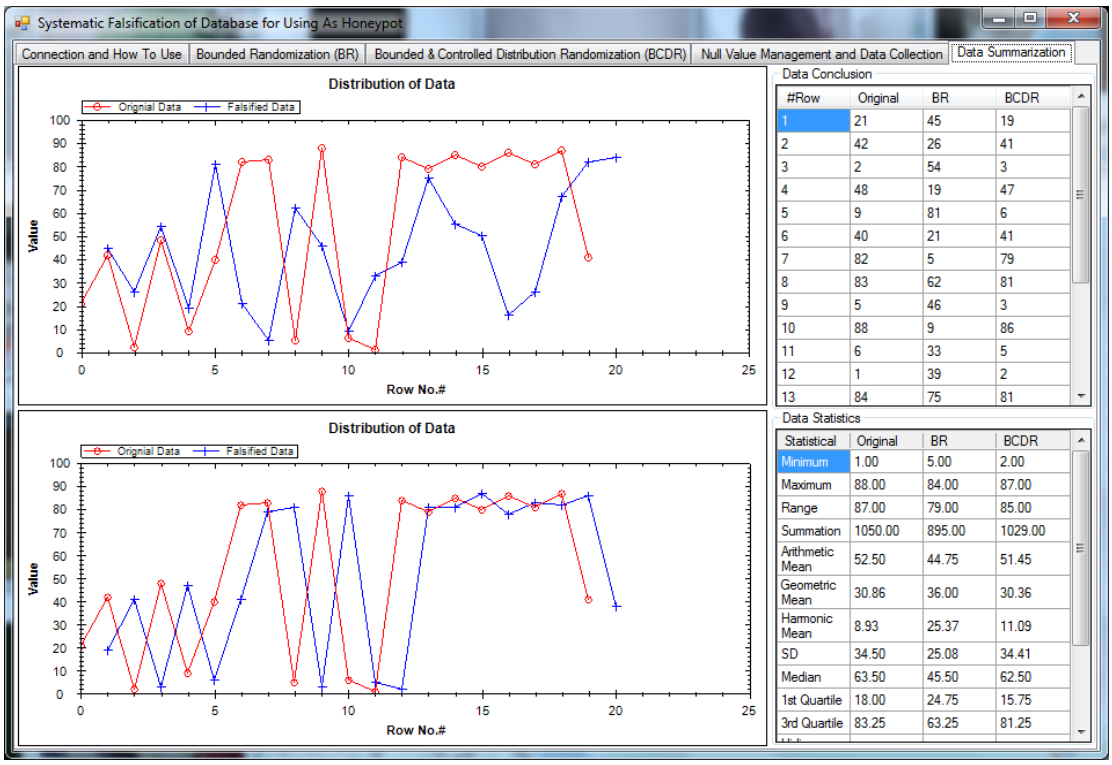
รูปที่ ก. 8 แสดงตัวอย่างของข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขต



รูปที่ ก.9 แสดงตัวอย่างข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยการควบคุมขอบเขตและการกระจายของข้อมูล



รูปที่ ก.10 แสดงตัวอย่างของข้อมูลที่ได้หลังจากการปรับแต่งจำนวนตัวเลขที่แสดงหลังจุดทศนิยม จำนวนค่าว่าง และการเลือกข้อมูลเพื่อบันทึก



รูปที่ ก.11 แสดงตัวอย่างของข้อมูลที่ได้โดยเปรียบเทียบข้อมูลระหว่างข้อมูลที่ได้จากการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการสุ่มค่าข้อมูลโดยควบคุมขอบเขตและการกระจาย

ประวัติผู้เขียนวิทยานิพนธ์

นายสิทธิเดช ท้วมพิบูลย์ เกิดเมื่อวันที่ 22 มกราคม พ.ศ. 2529 ที่จังหวัดสงขลา สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ จากภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ในปีการศึกษา 2550 และเข้าศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ปีการศึกษา 2551 งานวิจัยที่สนใจ ได้แก่ ความปลอดภัยของระบบคอมพิวเตอร์ การประมวลผลข้อมูล การทำเหมืองข้อมูล

Turnitin Originality Report
Systematic Falsification of Database for Using As Honeypot by Sithidech Tuampiboon
From Sithidech Tuampiboon (Thesis)

- Processed on 20-May-2012 13:54 ICT
- ID: 250494655
- Word Count: 26330

Similarity Index
5%
Similarity by Source

Internet Sources:
4%
Publications:
1%
Student Papers:
3%

sources:

1 < 1% match (student papers from 05/17/12)
Submitted to Chulalongkorn University on 2012-05-17

2 < 1% match (student papers from 05/16/12)
Submitted to Chulalongkorn University on 2012-05-16

3 < 1% match (student papers from 10/01/11)
Submitted to Chulalongkorn University on 2011-10-01

4 < 1% match (student papers from 05/01/12)
Submitted to Chulalongkorn University on 2012-05-01

5 < 1% match (Internet from 2/20/11)
http://esquero-me.exteen.com/

6 < 1% match (Internet from 5/18/09)
http://www.kmutt.ac.th/jgsm/2005/mambo/images/stories/JS%20klk.pdf

7 < 1% match (Internet from 5/13/09)
http://www.nii.ac.jp/graduate/thesis/pdf/kuruma_Dr_thesis.pdf

8 < 1% match (Internet from 10/25/10)
http://std2.co.eng.chula.ac.th/research-file/book/master/thesis.pdf

9 < 1% match (Internet from 10/27/10)
http://www.rubber.co.th/web/files/pdf/risk_book2553.pdf

10 < 1% match (student papers from 05/18/12)
Submitted to Chulalongkorn University on 2012-05-18

11 < 1% match (student papers from 10/05/11)
Submitted to Chulalongkorn University on 2011-10-05

12 < 1% match (Internet from 9/24/08)
http://www.cs.uq.oz.au/~xuel/ChenXia_ICDM04_WK.pdf

13 < 1% match (Internet from 9/24/10)
http://www.spsc.chula.ac.th/main2007/academic/pdf/2-2/09/2-07.pdf

14 < 1% match (Internet from 12/31/10)
http://www.thaihrhub.com/index.php/archives/research-view/research-686/

15 < 1% match (Internet from 6/26/10)
http://lib.cmu.ac.th/faculty/econ/Exer751409/2550/fullpaper_2550/Exer2550_52.pdf

16 < 1% match (Internet from 11/14/05)
http://astro.pudney.net.au/thesis/honeynet-thesis-pudney.pdf

17 < 1% match (Internet from 10/2/11)
http://www.upenlx.org/events/ikeit11/tech/full_papers/Nikiforakis.pdf

18 < 1% match (student papers from 02/27/11)
Submitted to University of Maryland, University College on 2011-02-27

19 < 1% match (Internet from 6/23/11)
http://www.gis.kmutnb.ac.th/thesisdata/4640580140.pdf

20 < 1% match (Internet from 5/6/10)
http://research.microsoft.com/en-us/um/people/thanis/phd-prize-2010dissertations/02-chrisli.pdf

21 < 1% match (Internet from 3/31/11)
http://www.cp.eng.chula.ac.th/~g47/ml/File1/example_proposal.doc

22 < 1% match (student papers from 03/08/11)
Submitted to Texas A&M University - Corpus Christi on 2011-03-08

23 < 1% match (Internet from 10/23/10)
http://library.cmu.ac.th/faculty/econ/Exer751409/2547/Exer47_no.4.pdf

24 < 1% match (Internet from 12/15/09)
http://portal.acm.org/citation.cfm?id=872771&jmp=abstract&coll=ACM&dl=ACM&CFID=27728484&CFTOKEN=90278860

25 < 1% match (Internet from 6/17/09)
http://pse-compliance.blogspot.com/2008_04_13_archive.html

26 < 1% match (publications)
Barbara Endicott-Popovsky, "Use of Deception to Improve Client Honeypot Detection of Drive-by-Download Attacks", Lecture Notes in Computer Science, 2009

27 < 1% match (student papers from 10/05/11)
Submitted to Chulalongkorn University on 2011-10-05

28 < 1% match (student papers from 10/27/09)
Submitted to Laureate Higher Education Group on 2009-10-27

29 < 1% match (Internet from 2/12/11)
http://www.ms.tku.ac.th/msllib/download/2552/43.pdf

30 < 1% match (Internet from 6/26/10)
http://lib.cmu.ac.th/faculty/econ/Exer751409/2550/fullpaper_2550/Exer2550_23.pdf

31 < 1% match (Internet from 9/1/10)
http://portal.acm.org/citation.cfm?doid=335191.335438

32 < 1% match (Internet from 2/12/11)
http://www.ms.tku.ac.th/msllib/download/2553/12.pdf

33 < 1% match (Internet from 4/26/11)
http://agenrs.faik.cvut.cz/cgi-bin/docarc/public.pl/document/210/main.pdf

34 < 1% match (student papers from 05/12/12)
Submitted to Chulalongkorn University on 2012-05-12

35 < 1% match (Internet from 4/18/11)
http://thesis.swu.ac.th/swuthesis/Mark/Kunyavee_P.pdf

36 < 1% match (publications)
L4KS V. S. Lakshmanan, "On disclosure risk analysis of anonymized itemsets in the presence of prior knowledge", ACM Transactions on Knowledge Discovery from Data, 10/01/2008

37 < 1% match (student papers from 07/30/09)
Submitted to Laureate Higher Education Group on 2009-07-30

38 < 1% match (Internet from 11/13/08)
http://www.ga.ukl.br/seminarios/CPE325/tutoriais/eugenia/honeynets.PDF

39 < 1% match (student papers from 10/12/11)
Submitted to Chulalongkorn University on 2011-10-12

40 < 1% match (student papers from 09/21/11)
Submitted to Chulalongkorn University on 2011-09-21

41 < 1% match (student papers from 05/13/12)
Submitted to Chulalongkorn University on 2012-05-13

42 < 1% match (student papers from 07/04/11)
Submitted to Naresuan University on 2011-07-04

43 < 1% match (Internet from 8/30/10)
http://www.chanpa.kku.ac.th/chanpa/Thesis/2543/word/plengsak.doc

44 < 1% match (Internet from 8/21/11)
http://www.nsit.ac.in/department/coeres.php

45 < 1% match (Internet from 5/20/10)
http://www.math.ucla.edu/~shargel/evolving.computer.scripts.pdf

46 < 1% match (student papers from 05/24/10)
Submitted to Laureate Higher Education Group on 2010-05-24

47 < 1% match (student papers from 10/13/11)
Submitted to Chulalongkorn University on 2011-10-13

48 < 1% match (Internet from 11/23/09)
http://www.businesshotel-kabe.co.jp/reserve/reserve.cgi

49 < 1% match (publications)
Submitted to GradeGuru

50 < 1% match (student papers from 05/12/12)
Submitted to Chulalongkorn University on 2012-05-12

51 < 1% match (student papers from 05/02/12)
Submitted to Chulalongkorn University on 2012-05-02

52 < 1% match (student papers from 00/00/00)
於2012-04-30提交至Chulalongkorn University

53 < 1% match (student papers from 04/11/12)
Submitted to Chulalongkorn University on 2012-04-11

54 < 1% match (student papers from 05/16/12)
Submitted to King Mongkut's University of Technology Thonburi on 2012-05-16

55 < 1% match (student papers from 07/01/05)
Submitted to CITY College, Affiliated Institute of the University of Sheffield on 2005-07-01

56 < 1% match (student papers from 05/15/07)
Submitted to Edith Cowan University on 2007-05-15

57 < 1% match (Internet from 6/26/10)
http://lib.cmu.ac.th/faculty/econ/Exer751409/2550/fulltpaper_2550/Exer2550_93.pdf

58 < 1% match (Internet from 4/22/10)
http://www.gis.kmutnb.ac.th/thesis-data/4740683408.pdf

59 < 1% match (Internet from 12/2/09)
http://dbis.ipd.uni-karlsruhe.de/download/SS07Graf_PrivatsphaereDataMining.pdf

60 < 1% match (publications)
A. Lafourcade, "Optimal sectionalization of a trellis", IEEE Transactions on Information Theory, 5/1996

61 < 1% match (student papers from 10/11/11)
Submitted to Chulalongkorn University on 2011-10-11

62 < 1% match (Internet from 10/29/10)
http://mirds.or.th/U-IRC/download/U-IRC_Handbook_Final.pdf

63 < 1% match (Internet from 10/31/11)
http://202.129.0.151/upload/090800/%E0%B8%84%E0%B8%B5%E0%B8%A2%E0%B8%A0%E0%B8%A2%E0%B8%E1%E0%B9%80%E0%B8%84%E0%B8%A3%E0%B8%B7%E0%B9%88%E0%B8%AD%E0%B8%87%E0%B8%A1%E0%B8%B7%E0%B8%AD%E0%B8%84%E0%B8%A7%E0%B8%B2%E0%B8%A1%E0%B9%80%E0%B8%A4%E0%B8%B7%E0%B9%88%E0%B8%AD%E0%B8%A1%E0%B8%B1%ED%B2%8%ED%B8%99.pdf

64 < 1% match (Internet from 8/27/10)
http://www.ru.ac.th/scholarship/STD_LOAN/resultyellow_1.asp?page=10

65 < 1% match (Internet from 11/11/10)
http://www.tlcp.co.th/upload_files/PDF/environmentMeeting1-6-2007-112.pdf

66 < 1% match (Internet from 4/16/11)
http://www.coursehero.com/file/2221841/ID-110/

67 < 1% match (Internet from 9/26/11)
http://www.sosaithanoen.ro/Thai_F2ooS.pdf.file

68 < 1% match (publications)
R. McGrew, "Experiences with Honeypot Systems: Development, Deployment, and Analysis", Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS 06), 2006

69 < 1% match (student papers from 10/09/11)
Submitted to Chulalongkorn University on 2011-10-09

70 < 1% match (student papers from 11/02/11)
Submitted to University of Fairfax on 2011-11-02

71 < 1% match (Internet from 6/26/10)
http://lib.cmu.ac.th/faculty/econ/Exer751409/2550/fulltpaper_2550/Exer2550_51.pdf

72 < 1% match (Internet from 3/18/11)
http://www.ckc.ac.th/master/project07.pdf

73 < 1% match (Internet from 10/25/11)
http://app.eng.ubu.ac.th/~edocs/t/20101203Thanarat62.pdf

74 < 1% match (Internet from 4/13/12)
http://www.kanchanaburi.go.th/library/pdf_research/chwa.pdf

75 < 1% match (Internet from 8/17/11)
http://softcreate.exteen.com/page/2

76 < 1% match (Internet from 6/27/10)
http://www.spt.co.th/jobs/?id=9

77 < 1% match (Internet from 12/31/10)
http://www.thaihrhub.com/index.php/archives/research-view/research-255/

78 < 1% match (Internet from 3/31/11)
http://www.it.kmitl.ac.th/th/publication/content/519

< 1% match (student papers from 10/13/11)

79 Submitted to Chulalongkorn University on 2011-10-13

80 < 1% match (student papers from 05/14/12)
Submitted to Chulalongkorn University on 2012-05-14

81 < 1% match (publications)
Liu Tian-Hua, "Core Functions Analysis and Example Deployment of Virtual Honeynet", 2011 First International Conference on Robot Vision and Signal Processing, 11/2011

82 < 1% match (publications)
Anand Gupta, "Opaqueness Characteristic of a Context Honeypot System", Information Security Journal A Global Perspective, 2010

paper text:

1 **2554** SYSTEMATIC FALSIFICATION **OF** DATABASE **FOR** USING AS HONEYPOT **Mr.** Sithidech Tuampiboon **A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of** Science **Program in** Computer Science **Department of** Computer **Engineering Faculty of Engineering Chulalongkorn University Academic Year 2011 Copyright of Chulalongkorn University**

2 . ..... **(.) .....** ..... **(.)** ..... **(.)** ..... **(**

.) : . (SYSTEMATIC FALSIFICATION OF DATABASE FOR USING AS HONEYPOT) . : .. , 95 . 3

4 . 2544 # # S170495021 : **MAJOR COMPUTER SCIENCE KEYWORDS:**

HONEYPOT / DATA PERTURBATION / PRIVACY-PRESERVATION / RELATIONAL DATABASE SITHIDECH TUAMPIBOON : SYSTEMATIC FALSIFICATION OF DATABASE FOR USING AS HONEYPOT. ADVISOR : YUNYONG TENG-AMNUAY,

53**Ph.D,** 95 **pp. The objective of this research is to propose**

the systematic falsification of database for using as Honeypot. This will normalize data to avoid the attacker's awareness and preserve the information privacy. This research proposes 3 main techniques. First, methodology for randomization of data including randomization of data by controlled boundary, randomization of data by controlled boundary and distribution of data, and data swapping. Then, methodology for modification data including values which are used as boundary and summary values. Finally, methodology for modifying data presentation including null value management, number of digits after decimal point management, and repetitive pattern of data management. Experimental results indicated that the proposed technique is effective in producing normalized data and in preserving confidentiality of original data.

8**Department Computer Engineering Student's Signature Field of Study Computer** Science **Advisor's Signature Academic Year** 2011 .

79

. .

62

71

40

80

3 **1 1.1 1 1.2** 3 **1.3** 4 **1.4**

52 **4** 1.5 **4** 1.6

105 2 6 2.1 6 2. 2 .1

(Honeypot) 6 2.2.2 (Honeynet) 11 2.2.3 (Data Perturbation) 14 2.2 15 2.2.1 (Honeyfile) 15 2.2.2 15 2.2.3 15 2.2.4 15 2.2.5 16 3 17 3.1 17 3.2 18 3.2.1 18 3.2.2 20 3.2.3 24 3.2.4 25 3.2.5 27 3.2.6 30 3.2.7 32 3.2.8 33 3.2.9 35 4 38 4.1 38 4.2 38 4.3 39 4.3.1 39 4.3.2 40 4.3.3

77

41 4.3.4 43 4.4 44 4.4.1 44 4.4.2 49 4.4.3 55 4.4.4 59 4.4.5 64 4.4.6 69 4.4.7 71 4.4.8 73

345 76 5.1 76 5.2 78 80 84

4185 95 2.1

13 2.2

32

14 3.1 23 3.2 31 3.3 36 4.1 40 4.2 41 4.3 42 4.4 r 43 4.5 45 4.6 4.5 47 4.7 49 4.8 51 4.9 4.8 53 4.10 55 4.11 57 4.12 60 4.13 x=10% 61 4.14 63 4.15 64 4.16 65 4.17 67 4.18 68 4.19 70 4.20 72 4.21 74 2.1 8 2.2 2 12 3.1 A = {48, 40, 54, 28,26, 34,32,44} 23 3.2 24 4.1 44 4.2 1 48 4.3 2 48 4.4 3 48 4.5 1 53 4.6 2 54 4.7 3 54 4.8 58 4.9 1 58 4.10 2 59 4.11 3 59 .1 Connection and How To Use 85 .2 Bounded Randomization (BR) 86 .3 Bounded & Controlled Distribution Randomization (BCDR) 87 .4 Null Value Management and Data Collection 88 .5 Data Summarization 90 .6 91 .7 92 .8 92 .9 93 .10 93 .11 94

27 **1 1.1**

3-4 Lance Spitzner (Honeypot) [1, 2, 3, 4] Honeyd [5] Specter [6] Honeywall CDROM ROO [7, 8] mwcollect [9] (Honeynet) [10, 11, 12] (Honeywall) () (File Server) (E-mail Server) (Web Server) (Database Administrator) 2008 5% [13, 14]

29

2

29 2 1)

200 2)

30 2 1)

2) Microsoft SQL Server 2005 (Database Management System – DBMS) (Integer) 1.2 3 1.3 1) 2) 3) 4) (Windows) 2008 R2 (MS SQL Server 2008 R2) (C#)

191.4 1)

2) 3)

214) 5) 6) 4 1.5

1.6

50

"SYSTEMATIC FALSIFICATION OF NUMERICAL DATABASE FOR USING AS HONEYPOT" . "The Seventh

78National Conference on Computing and Information Technology (NCCIT

2011)" 11-12 2554 5

42 2

2.1 2.1.1 (Honeypot) 2.1.1.1 (Passive Defense) [15] (Firewall) [16, 17] (IDS Intrusion Detection System) [18] Lance Spitzner " " 2 1) "" (Workstation) (File Server) (Mail Server) (Printer) (Router) 2) (Production Value) 2.1.1.2 1) (Network Device Hardware) : (Switch) 2) (Monitoring/logging Tools) : 3) (Management Workstation) : 4) (Alerting Mechanism) : 5) (Keystroke Logger) : 6) (Packet Analyzer Sniffer) : 7) (Data Backup) : 8) (Forensic Tools) : 9) (Research Resource) : 2.1 7 8 2.1 2.1.1.3 (Interaction Level) 2 1) (Low-Interaction Honeypot) mwcollect nepenthes honeytrap 2) (High-Interaction Honeypot) 2.1.1.4 1) (FP : False-Positive, FN : False- Negative) 2) (Early Detection) (Honeytoken) (Object) Administrator Administrator 9 Administrator 3) (New Threat Detection) 4) (Know Your Enemy) 5) (Honeypot As a Forensics Tools) 2.1.1.5 1) 2) (Web Server) 10 length lenght 3) 4) 2 (Gen‖) 2.1.2 (Honeynet) (Honeypot network) Microsoft SQL Server 2005 Windows XP Apache Web Server Ubuntu Domain Name Server Linux (Honeywall) 11 2 (Gen‖ Honeynet Generation II Honeynet) 2.2 12 2.2 2 2.2 3 1) (Production traffic) 2) (Honeywall Management traffic) 3) 2.2 3 4 1) (Data Control) DoS (Denial of Service) (South Florida Project) " 5-10 " 2) (Data Capture) 3) (Data Analysis) 13 2.1 Honeynet Project & Research Alliance (Honeywall CDROM Roo) 2.1.3 (Data Perturbation) [19, 20, 21, 22, 23] (Random Noise ε) A' = A + ε A' A ε 2.2

32

A A' ε 1 36 5 41 2 25 2 27 3 27 -9 18 4 33 20 53 5 31 15 46 6 37 27 64 14 2.2

102.2 2.2.1

(Honeyfile) Jim Yuill [24] password.txt 2.2.2 Neil C. Rowe [25] 2.2.3 (Honeypot Database) S. K. Gupta [26] OCHD (Obliviousness Characteristic of Honeypot Database) 2.2.4 M. Krishnamurty S.Rathindra Bias Corrected Correlated Noise (BCCN) Correlated Noise (CN) Independent Noise (IN) 15 2.2.5 B. Elisa [27] 4 1) (Privacy Level) 2) (Hiding Failure) 3) (Data Quality) 4) (Complexity) 16

72 3

(Relational Database) (Data Type) 2 (Numeric Data) (Alphanumeric Strings) (Cryptography) (Adding Noise) (Replacing) (Data Swapping) "" ( ) ( ) (Server Room) 3.1 3 3.1.1 3.1.2 3.1.3 3.2 18 3.2.1 300 3 "" [min, max] = { x | min ≤ x ≤ max} min max x 20 35 1 A A = {A1, A2, A3, …, AN} A'

7A' = (A' 1 , A' 2 , A' 3 , …, A'

N } Min Min = min{A1, A2, A3, …, AN} Min ≠ Max Max Max = max{A1, A2, A3, …, AN} Min ≠ Max ε N 1 Ax' = Ax + εx 2 1 (A x ') [Min, Max] Ax' = [Min, Max] Ax' = Ax + εx Ax + εx = [Min, Max] 3 εx Ax Min Max Ax + εx εx = [Min – Ax, Max – Ax] 4 3 εx 0 Ax εx εx = [Min – Ax, Max – Ax] – {0} 5 εx Ax A' = {A1 + ε1, A2 + ε2, A3 + ε3, …, AN + εN} A ε 4 ε A 5 A = {5, 6, 7, 8, 9} A = {5, 9} A 1 = 5 ε 1 ε1 = [Min – Ax, Max – Ax] – {0} ε 1 = [5 – 5, 9 – 5] – {0} ε 1 = [0, 4] – {0} ε 1 = {0, 1, 2, 3, 4} – {0} ε 1 = {1, 2, 3, 4} A1' = A1 + ε1 A1' - ε 1 = 1 A 1 ' = 5 + 1 = 6 19 - ε 1 = 2 A 1 ' = 5 + 2 = 7 - ε 1 = 3 A 1 ' = 5 + 3 = 8 - ε 1 = 4 A 1 ' = 5 + 4 = 9 A1' [5, 9] A 5 = 9 ε 5 ε5 = [Min – Ax, Max – Ax] – {0} ε 5 = [5 – 5, 9 – 9] – {0} ε 5 = [–4, 0] – {0} ε 5 = {–4, –3, –2, –1, 0} – {0} ε 5 = {–4, –3, –2, –1} A5' = A5 + ε5 A5' - ε 5 = –4 A 1 ' = 9 + (–4) = 5 - ε 5 = –3 A 1 ' = 9 + (–3) = 6 - ε 5 = –2 A 1 ' = 9 + (–2) = 7 - ε 5 = –1 A 1 ' = 9 + (–1) = 8 A5' [5, 9] 3.2.2 Agrawal, R. Srikant, R. 20 A A = {A1, A2, A3, …, AN} A'

7A' = (A' 1 , A' 2 , A' 3 , …, A'

N } ε N εx [Min – Ax, Max – Ax] – {0} εx ±x% Ax x εx 21 εx ±x% Ax Ax' = Ax + εx εx = {-x% Ax, +x% Ax] – {0} Ax Ax' = {Ax - x% Ax, Ax + x% Ax] – {Ax} Ax ± x% Ax - Ax ± x% Ax < Min Min Ax ± x% Ax - Ax ± x% Ax > Max Max Ax ± x% Ax - Ax ± x% Ax Ax ± x% Ax Ax ± x% Ax Ax' A' = {A1', A2', A3', …, AN'} A = {48, 40, 54, 28, 26, 34, 32, 44} x = 10%

73 3.1 3.1

22 23 60 50 40 30 20

63 1 2 3 4 5 6 7 8

3.1 A = {48, 40, 54, 28, 26, 34, 32, 44} 3.1 (x) Ax Ax - x% Ax Ax + x% Ax Ax' 1 48 43 53 45 35 2 40 36 44 41 50 54 3 54 49 59 52 29 4 28 26 25 31 31 38 26 5 26 23 29 28 44 6 34 31 37 31 46 7 32 29 35 35 26 8 44 40 48 47 49 3 A 3 + 10% A 3 = 59 54 4 A 4 - 10% A 4 = 25 26 5 A 5 - 10% A 5 = 23 26 24 A' A' = {45, 41, 52, 31, 28, 31, 35, 47} A' 3.2 60 50 40 30 20

64 1 2 3 4 5 6 7 8

3.2 3.2 () () () 3.2.3 (Data Swapping) 1 3 1150A 1150A 50,000 1150A 30,000 5,000 6,000

76 10,000 20,000

30,000 50,000 3.2.1 3.2.2 3.2.4 2 SELECT MIN(OrderPrice) AS SmallestOrderPrice FROM Orders 25 A A = {A1, A2, A3, …, AN} A'

7A' = (A' 1 , A' 2 , A' 3 , …, A'

N } Min Min ≠ Max Min = min{A1, A2, A3, …, AN} Max Min ≠ Max Max = max{A1, A2, A3, …, AN} Min' Min Min' ≠ Max' Max' Max Min' ≠ Max' N N A Min Max A = {A1, A2, A3, Min, Max, …, AN} Min Max Min' Max' A' = {A1, A2, A3, Min', Max', …, AN} 26 3.2.5 SELECT SUM(OrderPrice) AS OrderTotal FROM Orders 2 A A = {A1, A2, A3, …, AN} A'

7A' = (A' 1 , A' 2 , A' 3 , …, A'

N } sum A sum' sum sum ≠ sum' Δ N 3.2.5.1 1 A sum = ◆ =1 27 2 3 sum' / N 4 Δ 5 Δ A sum < sum'

60A' = (A1 +λ A3 +λ,…, AN +λ sum > sum' A' = {A1 =\ A2 =\

A3 –Δ 2 Δ Δ A = {48, 40, 54, 28, 26, 34, 32, 44} sum' = 500 1 A sum = 48 + 40 + 54 + 28 + 26 + 34 + 32 + 44 = 306 2 A 3 4 Δ 24.25 28 5 Δ A sum < sum' A' = {48 + 24.25, 40 + 24.25, 54 + 24.25, 28 + 24.25, 26 + 24.25, 34 + 24.25, 32 + 24.25, 44 + 24.25} A' = {72.25, 64.25, 78.25, 52.25, 50.25, 58.25, 56.25, 68.25} A' sum(A') = 72.25 + 64.25 + 78.25 + 52.25 + 50.25 + 58.25 + 56.25 + 68.25 = 500 3.2.5.2 1 (

Sum' < Sum) (Sum' > Sum) {6, 7, 8, 9, 10} Sum'= 60 Sum' > Sum {7, 8, 9, 10} 2 Sum{7, 8, 9, 10} = 34 3 Sum{6} = 6 29 4 3 |60 – 6| = 54 5 2 2 34/4 = 8.5 6 4 4 54/4 = 13.5 7 5 6 |8.5 – 13.5| = 5 8 7 (Sum' > Sum) (Sum' < Sum) {7 + 5, 8 + 5, 9 + 5, 10 + 5} = {12, 13, 14, 15) 9 1 8 {6, 12, 13, 14, 15} Sum{6, 12, 13, 14, 15} = 60 (6) 3.2.6 (Null)

65

2 4 30 3.2.6.1 3.2.6.2 3.2.6.3 3.2.6.4 3.2 3.3.5.1 3.3.5.2 3.3.5.3 3.3.5.4 1 21 30 30 30 30 NULL 2 42 41 41 41 NULL NULL 3 NULL 28 28 NULL 28 28 4 48 40 40 40 40 40 5 NULL 43 43 43 43 43 6 NULL 50 50 50 50 NULL 7 82 88 88 88 88 88 8 83 30 30 NULL 30 NULL 9 NULL 31 31 NULL NULL 31 10 88 25 25 NULL NULL NULL NULL 3.2 4 3 5 6 9 3.2.6.1 3.2.6.2 31 3.2.6.3 3 3.2.6.4 3.2.6.3 3 – 5 5 3.2.7 2 3 2 3 3.2.7.1 4.55 6.312 6.31 3.2.7.2 {5.123, 6.881, 9.013} 2 {5.12, 6.88, 9.01} 3.2.7.3 1-4 1 4 {5.12, 6.123, 7.9, 10.11} 32 3.2.8 0 0 {25650, 35550, 55320} 0 {25600, 35500, 55300} 1 0 () {5000, 45000, 600000} 600000 (6) 5000 45000 0 005000 045000 2 0 0 "0" 0 "X" (005000, 045000, 600000) 0 "0" 0 "00" 0 0 "000" {6, 5, 0} "-X000" "XXX0000" 33 3 "X" 2 "0" "X" "0" "0"

67

"X" "X" "X" {52000, 45000, 60000} "XX000" "X" "X" 2 3 {52, 45, 60} "000" 4 5 3 4 3 {52, 45, 60) 4 {50, 59, 47) "000" {50000, 59000, 47000} 6 "0" "X" "X0X00" "0" (10300, 20400, 50600) "X0X00" 3 (103, 204, 506} (425, 321, 221} 5 {42500, 32100, 22100} "0" "X" (40500, 30100, 20100} 34 3.2.9 1) 2) 3) 4) 5) 6) 7) 8) 9)

61 3.3 35 3.3

36 37 4 3 4.1

74

2 4.1.1 4.1.1.1 4.1.1.2 4.1.1.3 2 4.1.2 4.2 2 39 4.2.1 (Hardware) 1) (Windows 7) 2) (Intel Core 2 Duo 2.10Ghz) 3) 4 4.2.2 (Software) 4.2.2.1 (Programming Language) 1) 2) 4.2.2.2 (Database Application) 1) MS SQL Server 2008 R2 4.3 4.3.1 (Hiding Failure HF) Oliveira Zaiane [28] (Sensitive Information) 0

75

HF = # (') # () Rp RestrictedPattern D' D #R p (D') #R p (D) #R p (D) #R p (D')

15 4.1 4.1

Rp(D') 1 5 5 1 2 10 6 0 3 11 4 0 4 8 7 0 5 7 7 1 #R p (D') = 2 HF = 2/5 = 0.4 4.1 HF HF = 0.4 HF HF HF = 0 4.3.2 = # X 100 40 #Inbound N

43 4.2 4.2

#Inbound (D') 1 5 4 0 2 6 8 1 3 7 9 1 4 8 15 0 5 9 6 1 = 9 = 5 #Inbound (D') = 3 = (2 * 100)/5 = 60% 4.3.3 (Pearson Correlation Coefficient) 2 (Simple Correlation) r

35 $r = n\sum XY - \sum X\sum Y \; 2 \; 2 \; \{ n\sum X - (\sum X) \}, n\sum Y - (\sum Y) \; 2$

[ ] X, Y 1 2 n 41

23 4.3 4.3

r r = .50 1.00 r = -.50 -1.00 r = .30 .49 r = -.30 -.49 r = .10 .29 r = -.10 -.29 r = .00 (Scatter Diagram) 3 1. (Positive Correlations) 2. (Negative Correlations) 3. (Zero Correlations)

23 4.4 42 43 4.4

r r r = -1

11 () () r

= 0 r = 1

11 () ()

4.3.4 (Histogram) 4.1 44 4.4 4.1 1) 20 = 34 = 20 = 517 2) 10,000 = 18 = 36 = 250417 3) 10,000 = 4.00 = 0.58 = 27813.91 4.4.1

5

3 3

01 2 3

47 4.5 4.5

1 2 3 1 24 21 34 30 2 29 26 30 32 3 26 30 25 25 4 20 33 23 28 5 26 25 22 34 6 24 31 23 30 7 25 21 31 32 8 26 30 28 28 9 21 23 34 34 10 31 28 28 26 11 30 27 29 21 12 26 32 22 22 13 28 29 33 21 14 23 29 21 25 15 27 29 30 20 16 25 28 34 34 17 34 29 32 32 18 27 34 30 20 19 25 28 20 28 20 20 31 31 27 34 34 34 34 20 21 20 20 0 0 0 100 100 100 -0.0142 0.1768 -0.1812 45 4.5 0

6 4.6 46 47 4.6

4.5 1 2 3 4.6 48 4.2 1 4.3 2 4.4 3 4.2 4.3 4.4 3 1 2 3

15 4.7 4.7

49 1 0 100 -0.0397 2 0 100 -0.0265 3 0 100 -0.0407 1 0 100 -0.0068 2 0 100 -0.0121 3 0 100 -0.2304 4.7 0 100 4.4.2

5

3 3

9 1 x = 10%

2 x = 50% 3 x = 90%

57 4.8 50 4.8

1 2 3 1 24 22 22 32 2 29 30 27 22 3 26 25 33 27 4 20 22 30 31 5 26 28 27 27 6 24 23 23 32 7 25 23 27 29 8 26 24 22 22 9 21 22 20 29 10 31 34 33 27 11 30 32 27 34 12 26 24 27 24 13 28 30 29 33 14 23 21 20 20 15 27 28 25 30 16 25 23 31 27 17 34 32 26 26 18 27 25 28 25 19 25 26 23 23 20 20 22 30 28 34 34 33 34 20 21 20 20 0 0 0 100 100 100 0.8819 0.2344 -0.0674 51 4.8 0 x 1 x 10% x 50% x 90% x x x 0

58 4.9 52 4.9

4.8 1 53 2 3 4.5 1 54 4.6 2 4.7 3

54 4.5 4.6 4.7

1 2 4.5 4.6 3 4.7 3

9 - 1 x = 10% -

2 x = 50% - 3 x = 90%

6 4. 10 4. 10

55 1 0 100 0.8778 2 0 100 0.1279 3 0 100 -0.0406 1 0 100 0.9542 2 0 100 0.3871 3 0 100 0.0589 4.10 0 100 4.8 x x x 4.4.3

5

3 3

61 2 3

51 4. 11 56 4. 11

1 2 1 24 25 20 25 2 29 31 26 26 3 26 27 24 25 4 20 29 23 27 5 26 24 21 31 6 24 30 34 23 7 25 20 31 24 8 26 25 27 20 9 21 20 26 26 10 31 27 30 30 11 30 26 29 24 12 26 25 27 28 13 28 34 25 21 14 23 24 24 26 15 27 26 28 34 16 25 21 26 26 17 34 26 25 29 18 27 28 20 25 19 25 26 26 20 57 3 20 20 23 25 27 34 34 34 34 20 20 20 20 0 0 0 100 100 100 0.3321 0.1629 0.1629 4.11 0

5 3

4.8 4.9 4.10 4.11 58 4.8 4.9 1 59 4.10 2 4.11 3 4.4.4

5

3 4 - 1 = 36 , = 18 - 2 = 36 , = 22 - 3 = 32 , = 18 - 4 = 32 , = 22 2

19 4. 11 4. 12 4.

12

13 1 2 3 4

1 24 20 34 32 23 2 29 20 35 21 31 3 26 32 23 22 29 4 20 33 22 27 27 5 26 22 34 24 25 6 24 19 35 22 30 7 25 24 31 32 32 8 26 18 34 20 32 9 21 24 35 23 25 10 31 35 23 25 32 11 30 22 33 27 22 12 26 29 35 25 31 13 28 20 30 22 29 14 23 35 31 28 30 15 27 24 29 18 25 16 25 36 22 29 26 17 34 18 23 32 28 18 27 23 36 22 29 19 25 26 28 24 29 20 20 23 25 25 28 34 36 36 32 32 20 18 22 18 22 0 0 0 0 100 100 100 100 -0.2368 -0.0532 0.0258 0.1167 60 4.13 x=10%

13 1 2 3 4

1 24 26 22 23 25 2 29 31 30 31 30 3 26 25 27 28 24 4 20 18 22 18 22 5 26 27 27 28 25 6 24 23 22 25 26 7 25 27 24 26 23 8 26 28 27 25 28 9 21 23 23 19 22 10 31 30 29 28 28 11 30 33 32 28 28 12 26 27 28 25 24 13 28 30 29 29 26 14 23 22 24 22 25 15 27 29 26 29 28 16 25 24 27 26 26 17 34 36 31 32 32 18 27 25 28 28 19 25 27 24 24 26 20 20 19 22 19 22 34 33 32 32 32 20 19 22 18 22 0 0 0 0 100 100 100 100 0.9367 0.8821 0.9076 0.8672 61 4.11 4.12 0 100 4 - 1 = 40 , = 10 - 2 = 40 , = 25 - 3 = 27 , = 25 - 4 = 27 , = 10 62 4.14 -

14

1 40-10 0 100 -0.0046 2 40-25 0 100 0.0200 3 27-25 0 100 -0.0298 4 27-10 0 100 0.0029 X = 10% 1 33-17 0 100 0.8281 2 33-25 0 100 0.7505 3 27-25 0 100 0.6460 4 27-17 0 100 0.7795 4 - 1 = 4.50 , = 0.00 - 2 = 4.50 , = 2.00 - 3 = 3.50 , = 2.00 - 4 = 3.50 , = 0.00 63 4.15 -

14

1 4.50-0.00 0 100 -0.2482 2 4.50-2.50 0 100 -0.4613 3 3.50-2.00 0 100 -0.0180 4 3.50-0.00 0 100 0.0127 X = 10% 1 4.50-0.18 0 100 0.6738 2 4.49-2.00 0 100 0.5669 3 3.50-2.00 0 100 0.4098 4 3.50-0.39 0 100 0.5715

30 4.13 4.14

0 100 4.4.5 3 4 - 1 = 600 - 2 = 400 64 65 - 3 = 600 - 4 = 400 4.16

13 1 2 3 4

1 24 29 18 29 19 2 29 34 23 34 24 3 26 31 20 31 21 4 20 20 14 25 15 5 26 31 20 31 21 6 24 29 18 29 19 7 25 30 19 30 20 8 26 31 20 31 21 9 21 26 15 26 16 10 31 36 25 36 26 11 30 35 24 35 25 12 26 31 20 31 21 13 28 33 22 33 23 14 23 28 17 28 18 15 27 32 21 32 22 16 25 30 19 30 20 17 34 39 34 39 29 18 27 32 21 32 22 19 25 30 19 30 20 20 20 14 25 15 34 39 34 39 29 20 20 14 25 15 517 607 403 617 417 4.15 - 1 2 ( ) 3 0 100 - 3 4 ( ) 0 100 1 2 4 - 1 = 270000 - 2 = 250000 - 3 = 270000 - 4 = 250000 66 4.17 X = 10%

14

1 268194 0.07 81 -0.0914 2 240765 0.08 91 -0.1099 3 279669 0.08 73 -0.0919 4 249851 0 100 -0.0867 1 268965 0.15 85 0.7528 2 240536 0.16 94 0.7577 3 270359 0.16 85 0.7680 4 240134 0.19 94 0.7677 4 - 1 = 30000 - 2 = 25000 - 3 = 30000 - 4 = 25000 67 4.18 X = 10% 1 30000.11 0.0034 75 0.0067 2 25000.05 0.0025 87 0.0049 3 30000.14 0.0028 75 0.0044 4 25000.28 0.0027 87 -0.0084 1 29999.91 0.0058 92 0.9315 2 24999.72 0.0047 100 0.6294 3 30000.30 0.0051 92 0.6341 4 24999.80 0.0059 99 0.6326 4.17 4.18 0 75-100 68 4.4.6 - 1 - 2 - 3 = 10 - 4 5-10 69 4.19 1 2 3 4 1 21 21 21 21 21 2 42 42 42 42 NULL 3 NULL 48 48 NULL 48 4 48 9 9 9 9 5 9 NULL 82 NULL 82 6 NULL 83 83 NULL 83 7 82 88 88 NULL 88 8 83 NULL 6 NULL 6 9 NULL 1 1 1 1 10 88 79 79 79 NULL 11 6 NULL 85 85 85 12 1 NULL 80 80 NULL 13 NULL 81 81 NULL NULL 14 79 87 87 NULL 87 15 85 51 51 51 NULL 16 80 82 51 21 80 17 NULL 85 88 NULL 81 18 81 6 1 NULL 48 19 87 80 87 21 80 20 51 NULL 9 NULL NULL 4.19 70 4.4.7

9 - 1

2 - 2 1-4 71 72 4.20 1 2 1 9.790 9.79 9.8 2 1.702 1.70 1.7020 3 1.028 1.03 1.0280 4 7.343 7.34 7.3430 5 9.641 9.64 9.641 6 9.644 9.64 9.6440 7 5.824 5.82 5.824 8 8.414 8.41 8.414 9 2.882 2.88 2.882 10 3.925 3.93 3.925 11 2.879 2.88 2.879 12 1.390 1.39 1.390 13 9.549 9.55 9.549 14 0.243 0.24 0.2 15 2.824 2.82 2.8240 16 2.901 2.90 2.90 17 4.725 4.73 4.7 18 0.773 0.77 0.7730 19 1.883 1.88 1.883 20 5.402 5.40 5.4 4.20 4.4.8 - 1 - 2 73 4.21 1 2 1 20560 10290 20530 2 30650 10600 30680 3 10440 20180 10420 4 60660 10640 60640 5 10230 20730 10250 6 10560 50770 10530 7 20780 20750 20790 8 30870 50930 30840 9 40870 40150 40880 10 50610 20750 50640 11 50840 50830 50860 12 10880 20050 10900 13 10490 40410 10500 14 20220 10240 20210 15 30230 20330 30260 16 50410 20720 50420 17 50110 30280 50080 18 10780 40240 10790 19 20330 50090 20300 20 50110 40340 50100 60660 50930 60640 10230 10240 10250 0 0 100% 100% -0.03823 0.999999 74 4.21 2 "X0XX0" 75

395.1 5

3 1) 2) 3) 4 - 3 0 - - - 0 - 77 . 5.2 – – – (Natural Language Processing) [29] 78 – – 0 0 "XX9" 9 79 [1] Joho, D. Active Honeypots [Online]. 2004. Available

56from: http://www.ifi. uzh .ch/ archive/mastertheses/DA_Arbeiten_2004/Joho_Dieter.

pdf [2009, November 19]. [2]

23Mokube, I. and Adams, M., Honeypots: concepts, approaches, and challenges. Proceedings of the 45th annual southeast regional conference (ACM- SE'45),

2007. [3]

55Spitzner, L. Honeypots: Definitions and Value of Honeypots [Online]. 2001. Available from: http://www. securityfocus .com/

infocus/1492 [2009, November 13]. [4]

70Spitzner, L. Honeypots: Tracking Hackers. USA: Addison-Wesley, 2003. [5] Provos, N.

46Developments of the Honeyd Virtual Honeypot [Online]. Available from: http://www.honeyd.org

[2009, November 20]. [6]

68NETSEC. Specter: Intrusion Detection System [Online]. Available from: http://www.specter.com/

[2009, November 25]. [7]

25Honeynet Project & Research Alliance. Know Your Enemy: Honeywall CDROM Roo [Online] 2005. Available from: http:// old .honeynet.org/papers/cdrom/roo/index.html

[2009, November 8]. [8] The Honeynet Project. Honeywall CDROM

81[Online]. Available from: https://projects.honeynet.org/honeywall/

[2009, November 6]. 81 [9] mwcollect Alliance. Collaborative Malware Collection and Sensing [Online]. Available from: http://code.mwcollect.org/ [2009, November 26]. [10] The

38Honeynet Project. Know Your Enemy: Defining Virtual Honeynets [Online]. 2003. Available from: http:// old .honeynet.org/papers/virtual/

[2009, November 18]. [11] The Honeynet Project.

49Know Your Enemy: Honeynets [Online]. 2006. Available from: http:// old .honeynet.org/papers/honeynet/

[2009, November 12]. [12] The

45Honeynet Project. Know Your Enemy: Learning about Security Threats. 2 nd ed. USA: Addison-Wesley,

2004. [13] Chickowski, E. Why Your Database Are Vulnerable To Attack – And What You Can Do About It [Online]. Available From: http://www.darkreading.com/ tech-center/2/Database_Security.html [2011, August 5]. [14] Oltsik, J. Database at Risk. [Online] 2009. Available from: www.enterprisestrategygroup.com/2009/09/esg-research-brief- databases-at-risk/ [2011, August 5]. [15]

16Rong, C. and Yang, G. Honeypots in Blackhat Mode and its Implications. Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies

(PDCAT 2003), 2003. [16] Abie, H. An Overview of Firewall Technologies [Online]. 2000. Available from: http://www.nr.no/publications/FirewallTechnologies.pdf [2009, November 11]. 82 [17]

18Cheswick, W., R., Bellovin, S., M. and Rubin, A., D. Firewalls and Internet Security, Repelling the Wily Hacker. 2 nd ed. USA: Addison-Wesley,

2003. [18] Axelsson, S. Research in intrusion-detection systems: A survey [Online]. 1999. Available from:

66http://www.cs.uiuc.edu/class/fa05/cs591han/ papers/ axelssonSurvey99 .pdf

[2011, August 19]. [19]

20Agrawal, R., and Srikant, R. Privacy-Preserving Data Mining. Proceedings of the 2000 ACM SIGMOD international conference on Management of data,

2000. [20]

59Denning, D.E., Cryptography and Data Security. Newyork: Addison-Wesley, 1982.

[21] Lin, J., Liu and J.

31Privacy Preserving Itemset Mining Through Fake Transactions. Proceedings of the 2007 ACM symposium on Applied computing,

2007. [22]

36Muralidhar, K., and Sarathy, R. Security of Random Data Perturbation Methods. ACM Trans. Database Syst,

1999. [23] Yao, Y., Huang, L., Yang, W., Luo, Y., Jing, W., and Xu, W.

24Privacy-preserving Technology and Its Applications in Statistics Measurements. Proceedings of the 2nd international conference on Scalable information systems,

2007. [24]

17Yuill, J., Zappe, M., Denning, D., and Feer, F. Honeyfiles: Deceptive Files for Intrusion Detection. Proceedings from the Fifth Annual IEEE SMC,

2004. [25]

26Rowe, N., C. Measuring the Effectiveness of Honeypot Counter-Counterdeception. Proceedings of the 39th Annual Hawaii International Conference on System Sciences,

2006. [26]

82Gupta, S., K., Damor, R., Gupta, A., and Goyal, V.

OCHD:

44Preserving Obliviousness Characteristic of Honeypot Database. 13th International Conference on Management of Data,

2006. 83 Proceedings of [27]

33Bertino, E., Lin, D., and Jiang, W. A Survey of Quantification of Privacy Preserving Data Mining Algorithms.

28[Online] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1. 125.4396 &rep =rep1&type=pdf

[2011, August 5]. [28]

12Oliveira, S.R.M., Zaiane, O.R.. Privacy preserving frequent itemset mining.

Proceedings of IEEE icdm Workshop on Privacy, Security and Data Mining, 2002.

[29]

37**Wikipedia. Natural language processing [Online]. Available from: http://en.wikipedia.org/wiki/Natural_language_processing**

[2012, April 24]. 85 3 1. 1.1 Systematic Falsification of Database for Using As Honeypot .1 Connection and How To Use - Connection Settings Test Connection "Connection Succeeded" - Choose Table and Column - Column Information How to use 1.2 86 .2 Bounded Randomization (BR) - Min-Max-Sum Configuration (MMS Cont.) - Generate Graph ( ) - Reset - Value Comparison 1.3 87 .3 Bounded & Controlled Distribution Randomization (BCDR) - Min-Max-Sum Configuration (MMS Cont.) - Range ... to ... [x, y] 3 x {-1, -2, -3, -4, -5} y {1, 2, 3, 4, 5} - Generate Graph ( ) - Reset - Value Comparison 1.4 88 .4 Null Value Management and Data Collection - Number Of Decimal Management o Reserve Number of decimal o Self-specific number of decimal o Random number of decimal between - Null Value Management o Reserve number of NULL (Random Position) o Fill all NULL values with proper value o Specific number of NULL o Random number of NULL within range - Final Step – Save Falsified Data - Overall Information with Null Value 1.5 89 90 .5 Data Summarization - - Data Conclusion - Data Statistic 91 2.

48**.6 .7 .8 .9 .10 .11**

.6 92 .7 . 8 93 .9 .10 94 .11 95

69 22 .. 2529

4 2550

2551