

Совместный рабочий семинар ЭКЕ и ЕВРОСТАТа
по конфиденциальности статистической информации
(Скопье, бывшая югославская республика Македония,
14-16 марта 2001 г.)

Рабочий доклад №16

Тема II: Влияние новых технологических разработок в программном обеспечении, средствах связи и вычислительных процессах на SDC (Контроль за соблюдением конфиденциальности статистической информации)

СОВЕРШЕНСТВОВАНИЕ ПРОЦЕССА ПОДАВЛЕНИЯ ЯЧЕЕК В РАМКАХ КОНТРОЛЯ ЗА ОБЕСПЕЧЕНИЕМ КОНФИДЕНЦИАЛЬНОСТИ СТАТИСТИЧЕСКИХ ДАННЫХ

Представленная работа

Представлена Университетом Ла Лагуна, Тенерифе, Испания¹

Аннотация: В данной работе представлена хорошо известная методика подавления ячеек, упоминаемая здесь как *полное подавление ячеек*. Далее описывается и благоприятно сравнивается с полным подавлением ячеек новая методика, называемая *частичное подавление ячеек*. В заключение мы представляем новый *комбинированный метод*, объединяющий самые ценные характеристики обеих методик и обеспечивающий более эффективный способ подавления чувствительных ячеек любых таблиц в целях их защиты.

I. КЛАССИЧЕСКАЯ МЕТОДИКА ПОДАВЛЕНИЯ ЯЧЕЕК

1. В отношении массивов данных в табулярной форме (подобно двумерной таблице, показанной на рисунке 1) с чувствительными ячейками (подобно ячейке с номинальным значением 22, соответствующим виду деятельности II и региону C на рисунке 1) для обеспечения их конфиденциальности при публикации таблицы широко используется методика подавления ячеек. Суть этой методики состоит в пропуске (или замене «звездочкой») соответствующих ячеек, чтобы посторонний «взломщик» не мог вычислить конфиденциальную информацию. Таким образом, необходимо удалить по крайней мере сами чувствительные ячейки, которые называются *первичными подавлениями*. Однако обычно, ввиду наличия математической зависимости между ячейками табулярных данных (например, наличия маргинальных сумм строк и т.д.), первичных подавлений недостаточно и возникает необходимость в подавлении также и других ячеек. Такие дополнительные (не являющиеся чувствительными) подавленные ячейки называются *вторичными подавлениями*.

| | Регион А | Регион В | Регион С | Всего |
|----------------------|----------|----------|----------|-------|
| Вид деятельности I | 20 | 50 | 10 | 80 |
| Вид деятельности II | 8 | 19 | 22 | 49 |
| Вид деятельности III | 17 | 32 | 12 | 61 |
| ВСЕГО | 45 | 101 | 44 | 190 |

Рис.1. Инвестиции предприятий в миллионах евро.

¹ Подготовил Хуан Хосе Салазар Гонзалес.

2. Конечно, задача определения «самых лучших» вторичных подавлений с тем, чтобы:
- гарантировать защиту информации в первичных подавлениях и
 - свести к минимуму потери информации ввиду неопубликования содержания показателей ячеек,

является нелегкой и обычно именуется задачей подавления дополнительных или вторичных ячеек. В данной работе мы будем обозначать ее как задача ППЯ (полного подавления ячеек). Возможное решение задачи ППЯ на примере таблицы из рис.1 представлено схемой на рис.2.

| | Регион А | Регион В | Регион С | Всего |
|-----------------------------|----------|----------|----------|-------|
| Вид деятельности I | * | 50 | * | 80 |
| Вид деятельности II | * | 19 | * | 49 |
| Вид деятельности III | 17 | 32 | 12 | 61 |
| ВСЕГО | 45 | 101 | 44 | 190 |

Рис.2. Возможное решение данной задачи ППЯ.

3. На практике недостаточно защитить только точное значение с помощью первичного подавления и «звездочки», - статистические учреждения также стремятся обеспечить неопределенность интервалов некоторых значений. Чтобы быть более точными в определении задачи ППЯ, введем некоторые основные концепции:

- **Внешние границы:** для каждой ячейки k мы называем *нижней внешней границей* lb_k минимальное ее значение, заранее известное взломщику, если данная ячейка будет подавлена. Обычно оно равно 0, однако на практике это может быть и другое число. Для каждой ячейки k мы называем *верхней внешней границей* ub_k максимальное ее значение, заранее известное взломщику, если данная ячейка будет подавлена. Обычно это очень большое число, но на практике может быть не столь большим. Обе внешние границы представляют собой априорную информацию, известную постороннему взломщику относительно подавленного значения, и поэтому они являются вводными параметрами для задачи ППЯ.
- **Конгруэнтная таблица:** при данной схеме подавления (аналогичной показанной на рис.2) и определенных внешних границах каждой подавленной ячейки *конгруэнтная таблица* представляет собой набор значений, который посторонний взломщик расценивает как исходную таблицу. Следовательно, конгруэнтная таблица совпадает с опубликованными значениями без подавления ячеек, удовлетворяет всем математическим уравнениям для таблиц и содержит конгруэнтные значения с внешними границами, которые соответствуют подавленным ячейкам.
- **Требования по уровню подавления:** при чувствительной ячейке k и трех вводных параметрах LPL_k , UPL_k и SPL_k конкретная схема подавления защищает ячейку k если и только если взломщик (зная только схему подавления и внешние границы подавленных значений) не может классифицировать как недопустимое, что:
 - в конгруэнтной таблице значение ячейки k меньше или равно номинальному значению минус LPL_k ;
 - в конгруэнтной таблице значение ячейки k больше или равно номинальному значению минус UPL_k ;
 - в двух конгруэнтных таблицах разность между значениями ячеек k больше или равна SPL_k .

Параметры LPL_k , UPL_k и SPL_k называются *нижним, верхним и скользящим уровнями защиты* соответственно. Обычно статистические учреждения выражают их в процентах от номинальных значений (например, 20%, 30% и 60% соответственно), но могут быть выбраны и другие числа. Естественно, чем больше эти параметры, тем больше потеря информации при обеспечении конфиденциальности данных.

- **Потери информации:** всякое подавление ячейки в схеме подавления влечет потерю информации. Для каждой ячейки k существует соответствующий параметр w_k , известный как *цена подавления* при подавлении ячейки k , а общая *потеря информации* для данной схемы подавления равна сумме стоимости всех подавлений.

4. Задача ППЯ подразумевает поиск такой комбинации подавлений, которая бы удовлетворяла условиям трехуровневой защиты всех чувствительных данных в таблице с минимальной потерей информации. Следовательно, эта задача является задачей *оптимизации*.

5. Задача ППЯ широко изучалась в различных работах (смотри, к примеру, Уилленбург и Де Ваал (1996)). Фишетти и Салазар (1999) предлагают новый алгоритм на основе современных методов математического программирования и решают задачи для двухмерных таблиц с 500 строками и 500 столбцами за несколько минут на самом обычном персональном компьютере. Фишетти и Салазар (2000) также расширяют свое решение для других k -мерных таблиц при $k > 2$, иерархических таблиц, связанных таблиц и т.д., получая при этом очень интересные результаты вычислений.

II. МЕТОДИКА ЧАСТИЧНОГО ПОДАВЛЕНИЯ ЯЧЕЕК

6. В классической методике полного подавления ячеек, даже если схема подавлений заключается в опубликовании или подавлении каждой ячейки таблицы, с точки зрения взломщика итоговая таблица представляет собой набор интервалов. Действительно, исходя из схемы подавлений и внешних границ взломщик может рассчитать *вероятные границы* каждого недостающего значения. Например, взломщик, рассматривающий схему подавления на рис.2 и знающий, что внешние границы отсутствующих значений лежат в интервале от 0 до бесконечности (т.е. значения, замещенные «звездочками», не являются отрицательными), знает также, что значения конгруэнтных таблиц находятся в интервалах, указанных на рисунке 3.

| | Регион А | Регион В | Регион С | Всего |
|----------------------|----------|----------|----------|-------|
| Вид деятельности I | [0...28] | 50 | [2...30] | 80 |
| Вид деятельности II | [0...28] | 19 | [2...30] | 49 |
| Вид деятельности III | 17 | 32 | 12 | 61 |
| ВСЕГО | 45 | 101 | 44 | 190 |

Рис.3. Вероятные границы значений для схемы подавления из рис.2.

7. Действительно, из таблицы на рис.2 и того факта, что отсутствующие значения не являются отрицательными, взломщик определит, что (к примеру) 1 не является возможным отсутствующим значением ячейки «Вид деятельности II – Регион С», даже если эта ячейка подавлена полностью. Для расчета вероятных границ даже для очень сложных таблиц наподобие связанных, взломщику нужен всего лишь самый стандартный и общедоступный оптимизатор линейного программирования. Сегодня практически все программное обеспечение для обработки данных (например, *MS Excel*, *Maple*, *Mathematica*, и т.д.) имеет оптимизаторы линейного программирования. Кроме того, в Интернете есть ряд общедоступных средств для решения задач по вычислению вероятных интервалов каких-либо параметров. Следовательно, классическую методику полного подавления ячеек можно рассматривать как метод публикования интервалов вместо конкретных данных типа «да-нет». Это, конечно, является общеизвестным наблюдением, известным как результат *аудиторской* фазы.

8. Разработка новой методологии мотивируется следующим вопросом. Если задача оптимизации при классическом (полном) подавлении ячеек предусматривает поиск оптимальных вариантов подставленных интервалов вместо точных исходных значений, то почему бы не найти также предельные значения этих интервалов?

9. Если целевая функция подавления ячейки сводится к минимизированию потери информации, то дополнительные расчеты предельных значений вероятностных интервалов дают возможность при оптимизации найти схемы подавления с меньшими потерями информации и в тоже время гарантировать тот же уровень защиты. К примеру, если статистическому учреждению требуются уровни защиты $LPL=2$, $UPL=4$ и $SRL=0$ для чувствительной ячейки из рис.1, то схема подавления на

рис.2 может являться схемой с минимальной потерей информации при использовании классического метода подавления ячеек с учетом определенных внешних границ и цены подавления. Но применяя новую методику (*частичного подавления ячеек*) возможны другие схемы подавления, удовлетворяющие требуемым уровням защиты и с меньшими потерями информации. В примере на рис.4 показана возможная схема, удовлетворяющая требуемым уровням защиты и с меньшими потерями информации по сравнению со схемой подавления на рис.3. Таким образом, первое преимущество новой методики по сравнению с классической заключается в том, что частичное подавление уменьшает потерю информации.

| | Регион А | Регион В | Регион С | Всего |
|-----------------------------|-----------|----------|-----------|-------|
| Вид деятельности I | [18...24] | 50 | [6...12] | 80 |
| Вид деятельности II | [4...10] | 19 | [20...26] | 49 |
| Вид деятельности III | 17 | 32 | 12 | 61 |
| ВСЕГО | 45 | 101 | 44 | 190 |

Рис.4. Возможное решение задачи ЧПЯ.

10. Второе очевидное преимущество использования новой методики подавления состоит в том, что «аудиторная фаза» становится лишней. Действительно, после решения задачи ППЯ необходимо также вычислить аудиторную фазу, чтобы рассчитать вероятностные границы подавлений. В новом методе эта дополнительная работа ненужна, поскольку вероятностные границы вычисляются автоматически при решении самой задачи ЧПЯ.

11. Судя по всему, проблема оптимизации при новой методике (задача ЧПЯ) является более сложной, чем при старой (задача ППЯ), но это не так! Сравнивая обе проблемы по оптимизации с помощью стандартных методов теории алгоритмов, задача ЧПЯ представляется намного легче задачи ППЯ. Если быть более точным, задача ППЯ является *NP*-сложной в строгом смысле этого слова, а задача ЧПЯ является полиномиально решаемой. Следовательно, третье преимущество использования новой методики в том, что проблема оптимизации (теоретически) становится намного легче.

12. Конечно, можно думать, что это последнее преимущество основывается только на теоретическом анализе, а на практике может быть очень трудно применить эффективный алгоритм решения задачи ЧПЯ. Но это не так, и подробности математического программирования изложены в работе Фишетти и Салазара (1998). Фактически, результаты вычислений из этой работы, приведенные в нижеследующей таблице, были получены на ПК Pentium 133 Mhz, ОЗУ 32 Мбит в среде *MS-Windows 95*.

| <i>type</i> | <i>cells</i> | <i>link</i> | <i>sensi</i> | <i>levels</i> | <i>sup</i> | <i>loss</i> | <i>Time</i> | <i>sup'</i> | <i>loss'</i> | <i>time'</i> | <i>time''</i> |
|-------------|--------------|-------------|--------------|---------------|------------|-------------|-------------|-------------|--------------|--------------|---------------|
| 41x31 | 1271 | 72 | 3 | 6 | 9 | 153 | 4.78 | 7 | 486 | 367.62 | 0.07 |
| 183x61 | 11163 | 244 | 2467 | 4934 | 6 | 140297 | 17.28 | 3 | 281398 | 11.42 | 667.06 |
| 359x46 | 16514 | 405 | 4923 | 9846 | 85 | 9357 | 77.67 | 38 | 19099 | 116.93 | 2486.07 |

13. Первые пять столбцов описывают характеристики трех реальных примеров: где *type* – внутренняя структура каждого образца таблицы, *cells* - количество ячеек, не подлежащих публикации, *link* - количество математических уравнений между ячейками таблицы, *sensi* - количество чувствительных ячеек, а *levels* – количество требуемых ненулевых уровней защиты. Следующие три столбца относятся к характеристикам задачи ЧПЯ: *sup* - количество ячеек с интервалами при оптимальном решении, *loss* - потери информации, *Time* - время вычисления в секундах на ПК Pentium 133. Последние четыре столбца описывают задачу ППЯ: *sup'* - количество отсутствующих значений при оптимальной схеме подавления, *loss'*- потери информации при оптимальной схеме подавления, *time'*- время вычисления в секундах для решения задачи ППЯ, а *time''*- время вычислений для аудиторской фазы.

14. Фундаментальная гипотеза в пользу предпочтения частичного подавления ячеек вместо классического подавления по схеме на рис.4 заключается в том, что при этом публикуется больше информации, нежели при схеме на рис.3 (которая эквивалентна схеме на рис.2 с точки зрения взломщика). Это объясняется тем, что более узкие интервалы дают более точную информацию.

15. Важное наблюдение заключается в том, что потеря информации при обеих методиках имеет разное значение. В то время, как в задаче ППЯ цена подавления w_k определяется ответом «да» или «нет» в зависимости от опубликования или подавления ячейки k , в задаче ЧПЯ потеря информации пропорциональна ширине публикуемого интервала. Цена потери единицы информации ниже или сверх номинального значения может быть различной, если это целесообразно для статистического учреждения. Другими словами, в задаче ППЯ статистическое учреждение должно принимать во внимание фиксированную цену подавления w_k для каждой непубликуемой ячейки k независимо от вероятностного интервала подавленных значений в итоговой схеме подавления. Например, ячейка с ценой подавления равной 100 прибавит к сумме потерянной информации именно 100, если эта ячейка не будет опубликована, независимо от вероятностного интервала подавленного значения в итоговой схеме подавления. В задаче ЧПЯ статистическое учреждение может рассматривать возможность потери информации пропорционально ширине публикуемых интервалов. Например, статистическое учреждение может установить потерю (скажем) 5 единиц информации для каждой единицы в вероятностном интервале ниже номинального значения и (скажем) 2 единиц для каждой единицы в вероятностном интервале сверх номинального значения. Тогда, ячейка с исходным значением 300, замещенная интервалом [290...320] подразумевает потерю информации равную 90, а при ее замещении интервалом [295...310] потеря информации составит 45. В целом в задаче ЧПЯ каждая ячейка k должна иметь два вводных параметра: w_k^- и w_k^+ .

III. СОВЕРШЕНСТВОВАНИЕ КЛАССИЧЕСКОЙ МЕТОДИКИ ПОДАВЛЕНИЯ ЯЧЕЕК

16. Из вышеописанных вычислительных экспериментов видно, что обычно частичное подавление ячеек имеет тенденцию замещать больше значений с помощью интервалов, чем замещается «звездочками» при использовании метода классического подавления ячеек. Замещение значения происходит даже при очень узком интервале, но это помогает уменьшить потерю информации. Чтобы устранить эту трудность очень просто скомбинировать обе методики и получить новый метод, предоставляющий статистическому учреждению также возможность дополнительно контролировать количество замещений, минимальную длину интервалов и т.д.

17. Действительно, математическая модель решения задачи ППЯ требует ряда бинарных переменных, одну переменную x_k для каждой потенциально подавляемой ячейки k :

$$x_k = \begin{cases} \text{если необходимо подавить ячейку } k \\ = 0 \text{ в остальных случаях} \end{cases}$$

Другими словами, подлежащая публикации схема будет содержать исходное значение ячейки k если $x_k = 0$, и замещать исходное значение «звездочкой» при $x_k = 1$. Тогда модель целочисленного программирования будет такова

$$\begin{aligned} &\text{минимизирование} \quad \sum_k w_k x_k \\ &\text{при условии:} \quad x_k \in \{0,1\} \text{ для всех ячеек } k \end{aligned}$$

плюс ряд линейных ограничений для установки необходимых уровней защиты для всех чувствительных ячеек. Идея записи этих требований в качестве линейных ограничителей

основывается на *теории дуальности* из линейного программирования и мы отсылаем читателя к работе Фишетти и Салазара (2000), где подробно отражена математическая сторона вопроса.

18. Точно также и для математической модели решения задачи ЧПЯ требуется ряд не отрицательных непрерывных переменных, две переменные z_k^- и z_k^+ для каждой потенциально замещаемой ячейки k :

z_k^- обозначает уменьшение относительно номинального значения
 z_k^+ обозначает увеличение относительно номинального значения

Другими словами, подлежащая опубликованию схема будет содержать $[a_k - z_k^- \dots a_k + z_k^+]$ вместо номинального значения a_k . Конечно, когда $z_k^- = z_k^+$, публикуется исходное значение ячейки k . Тогда схема целочисленного программирования будет такова

минимизирование $\sum_k (w_k^- z_k^- + w_k^+ z_k^+)$
 при условии: $z_k^- \geq 0$ для всех ячеек k ,
 $z_k^+ \geq 0$ для всех ячеек k ,

плюс ряд линейных ограничений для установки необходимых уровней защиты для всех чувствительных ячеек. Идея записи этих требований в качестве линейных ограничителей основывается на *теории дуальности* из линейного программирования и мы отсылаем читателя к работе Фишетти и Салазара (1998), где подробно отражена математическая сторона вопроса.

19. Для описания проблемы оптимизации при комбинированной методике подавления ячеек посредством объединения двух методов нам необходимо только добавить бинарные переменные в задачу ЧПЯ со следующими значениями:

$x_k = 1$ если интервалом надо заместить ячейку k
 $= 0$ в остальных случаях,

плюс следующие ограничители:

$z_k^- \leq (a_k - lb_k) x_k$ для всех ячеек k
 $z_k^+ \leq (ub_k - a_k) x_k$ для всех ячеек k

где a_k - номинальное значение, lb_k - нижняя внешняя граница, а ub_k - верхняя внешняя граница ячейки k . Эти ограничительные условия обеспечивают $x_k = 1$ при положительных значениях z_k^- или z_k^+ , т.е. когда точное значение a_k ячейки k предпочитается не публиковать.

20. Конечно, в новой математической модели имеются как непрерывные, так и целочисленные переменные, поэтому она относится к смешанному целочисленному математическому программированию и, следовательно, проблема оптимизации классифицируется как проблема *NP*-сложности в теории сложных алгоритмов. Таким образом, новая *комбинированная методика* не имеет преимуществ полиномиально решаемых проблем оптимизации в отличие от метода частичного подавления ячеек. Тем не менее, можно использовать алгоритм ветвления и отсечения для поиска эвристических и оптимальных решений для случаев набора данных среднего размера и,

следовательно, предлагаемую методику можно использовать. Основная концепция этого алгоритма аналогична алгоритму, предложенному Фишетти и Салазаром (2000) для решения задач ППЯ.

21. Что касается остальных двух преимуществ частичного подавления ячеек, которые отсутствуют при классическом подавлении ячеек, то в комбинированной методике оба преимущества сохраняются. Фактически, потеря информации может быть меньше и в аудиторной фазе нет необходимости, так как при использовании новой методики также вычисляются предельные значения вероятностных интервалов.

22. Более того, как отмечалось ранее, новая комбинированная методика позволяет также контролировать количество непубликуемых ячеек, поскольку математическая модель может включать неравенство

$$\sum_k x_k \leq \text{МАКС_ИНТЕРВАЛЫ}$$

для данного вводимого параметра МАКС_ИНТЕРВАЛЫ, задающего максимальное количество соответствующих интервалов, допускаемых нами в окончательной схеме подавления. Также можно заранее определить минимальную ширину подлежащих интервалов, рассматривая такие ограничители как:

$$z_k^- + z_k^+ \geq \text{МИН_ШИРИНА} \cdot x_k \quad \text{для всех ячеек } k$$

для данного вводного параметра МИН_ШИРИНА, представляющего минимальную ширину интервала в публикуемой схеме подавления, определяемую статистическим учреждением. И, наконец, можно контролировать и другие параметры окончательной схемы подавления, если их можно записать в качестве математических ограничителей для вышеуказанных переменных.

IV. ЗАКЛЮЧЕНИЕ

23. Мы представили три методики защиты чувствительных данных перед публикацией статистических таблиц. Первая методика – классическое подавление ячеек, называемое «полным подавлением ячеек» и заключающееся в отборе ячеек, которые можно заместить недостающими значениями (или «звездочками»). Вторая – это новая методика, называемая «частичным подавлением ячеек», при которой номинальные значения некоторых ячеек можно заменять интервалами. Третья методика объединяет преимущества двух предыдущих и является комбинированным способом подавления. Рассмотрены проблемы оптимизации, характерные для этих трех методик.

Список литературы

М. Фишетти, Х.Х. Салазар, «Частичное подавление ячеек: новая методика контроля за соблюдением конфиденциальности статистических данных», *рабочий доклад*, Университет Ла Лагуна, 1998.

М. Фишетти, Х.Х. Салазар, «Модели и алгоритмы для задачи подавления двухмерных ячеек при контроле за соблюдением конфиденциальности статистических данных», *Математическое программирование*, 84 (1999 г.), 283- 312.

М. Фишетти, Х.Х. Салазар, «Модели и алгоритмы для оптимизации подавления ячеек в табулярных данных с помощью линейных ограничителей», *Журнал Американской статистической ассоциации*, 451 (2000 г.).

Л. Уилленборг, Т. Де Ваал, «Контроль за соблюдением конфиденциальности статистических данных на практике», Конспекты лекций по статистике, Вып. 111, «Спрингер Верлаг», Нью-Йорк, 1996 г.