IDFinder: Data Visualization for Checking Re-identifiability in Microdata

Hyunmo Kang

Human Computer Interaction Lab.,
Department of Computer Science,
University of Maryland at College Park
kang@cs.umd.edu

Hyunmo Kang, Sam Hawala, Laura Zayatz

Statistical Research Division,
U.S. Census Bureau,
US Department of Commerce, United States
{hyunmo.kang, sam.hawala, laura.zayatz}
@census.gov

Abstract

This paper introduces a novel user interface, IDFinder, which is specifically designed to facilitate the disclosure avoidance process on microdata files. IDFinder is designed based on the well-known visual seeking mantra, “Overview first, Zoom and filter, and Details on demand”. Direct data manipulation and dynamic query techniques implemented in IDFinder provide users rapid, incremental and reversible operations, which are critical for disclosure avoidance tasks. Multiple tightly coupled data viewers are used to represent the different data hierarchies in microdata. In addition, time series viewers, which are also tightly coupled with other data viewers, visualize the change of attribute values over time and enable users to observe the attribute values in each data hierarchy at the specified time. The usability study with a small group of disclosure avoidance researchers led to the refined designs of IDFinder, and it also revealed benefits, scalability issues, and applicability to other tasks.

1 Introduction

One of the major functions of a federal statistical agency is to collect data from various sources (e.g. individuals, households, farms, businesses, and governmental bodies), process them and disseminate the results to the public in a variety of statistical forms. In the past, this dissemination was mostly in statistical and tabular form. With the advent of the computer age in the early 1960s, agencies also started releasing microdata files in an effort to meet the needs of researchers who required specialized tabulations (Ruggles, 2000). In a publicly released microdata file, each record contains a set of variables that pertain to a single respondent and are related to that respondent’s reported values. However, there are no identifiers on the file, so individual data items cannot be uniquely and easily associated with a particular respondent (Federal Committee on Statistical Methodology, 1994).

Nevertheless, the identity of an individual in a microdata file might be revealed by matching a respondent’s characteristics to a public or private data source that includes names. For example, there might be some sort of database containing name, age, sex and marital status for a subset of individuals in a locality. By searching the anonymized microdata file for persons with the same combination of characteristics, it might be possible to guess the identity of a respondent. This would result in what analysts have termed *re-identification disclosure* or *inferential disclosure*. The large amount of information easily accessible today and the increased computational power available to the attackers make such linking and matching attacks a serious concern (Samarati, 2001), (Federal Committee on Statistical Methodology 2002; Rasinski & Wright, 2000). All agencies which provide public use microdata files are using various methods for the protection of these microdata files. To reduce the potential for disclosure, virtually all such files include data from only a sample of the population, do not include obvious identifiers, limit geographic detail, and limit the number of variables in the file. Furthermore, to disguise high visibility variables, agencies also use additional mathematical methods which include top or bottom coding, recoding into intervals or rounding, adding or multiplying by random numbers (also called noise addition), swapping and rank swapping (sometimes called switching), blanking out selected variables and imputing for them, aggregating across small groups of respondents and replacing one individual's reported value with the average (also called blurring) (Federal Committee on Statistical Methodology, 1994; Hawala 2003; Hawala, Zayatz, & Rowland, 2003). All disclosure avoidance methods result in some loss of information, and sometimes the publicly available data may not be adequate for certain statistical studies. However, the intention is to provide as much data as possible without revealing individually identifiable data.

Although microdata files are pre-processed before they are released to the public, there is no definite answer for the question, "Can respondents in a microdata file still be re-identified using publicly available information sources?" This is because there are no accepted measures of disclosure risk for a microdata file. In other words, there is no "standard" which can be applied to assure that protection is adequate. This is also true because we can never be familiar with all publicly available datasets or know which sets will be made available in the future. Every re-identifiability checking process so far has been done locally (focusing only on a small part of microdata) and manually (e.g. writing SAS codes). It is a time-consuming and trial-and-error procedure. A typical procedure is composed of four steps: finding records in the microdata file that may be at risk of re-identification (henceforth called "possible targets"). These are often unique with respect to a few attribute values which makes them stand out from the rest of records in the microdata file; finding "possible suspects" (data records that have similar attribute values as those of the possible targets) using public data sources; linking and matching; and finally, masking the records if necessary. It has long been recognized that "finding possible targets" is the one step that is the most subjective and unsystematic of the four steps. The analyst has to guess at what attributes and what data an attacker might have available and might link to the microdata file to re-identify targets. To always make the correct guess the analyst would have to know all relevant data that are available, data that are being collected, and data that will be collected and made available in the future. This of course is impossible.

In this project, we designed and developed a prototype re-identifiability checking system, IDFinder, to facilitate the first step of this disclosure avoidance process and make it less dependent on the analyst. This is done through an information visualization approach. The main goal of this project is to enable the users, who are the employees of federal statistical agencies working on disclosure avoidance research, to visually identify any microdata records that are at risk of disclosure so that they can be masked before the microdata files are released to the public. The design of IDFinder is based on the well-known visual seeking mantra, "Overview first, Zoom and filter, and Details on demand" (Shneiderman & Plaisant, 2004). It also uses the direct data manipulation and dynamic query techniques (Card, Mackinlay, & Shneiderman, 1999) to provide users rapid, incremental and reversible operations, which are critical for disclosure avoidance tasks. IDFinder is designed to have multiple tightly coupled viewers to represent the different data hierarchies in microdata. In addition, it provides time-series viewers, which are also tightly coupled with other data viewers, so that users can select a time to see the corresponding values in each hierarchy as well as the change of values over time. The design of IDFinder focuses on three major issues. The first issue is about the scale of microdata. Microdata files usually contain hundreds of variables and hundreds of thousands of records. It is hard to represent more than a million items in a general display monitor without occlusion and to visualize more than 6 variables in 2D space simultaneously with limited visual cues such as spatial xy location, color, shape, size, and so on. Second, most microdata files have a hierarchy structure. For example, data from the Survey of Income and Program Participation (SIPP) is composed of 5-level hierarchies: household; family and subfamily; person; wave; and reference month. Since the hierarchy information is often used for expanding or reducing the search space, how to visualize multiple data hierarchies simultaneously, as well as how to link those hierarchies through visualizations, is an important issue. Finally, some microdata files contain data from repeated interviews. This is called time-series data. The change of attribute values over time is often used as a key for finding possible targets. Therefore, there needs to be a way to visualize and coordinate both the change of attribute values over time and the values at the specified time.

This paper starts with the data analysis of microdata and then introduces a novel user interface, IDFinder. An overview of the innovative features of IDFinder is followed by a usability study to validate the effectiveness of the system, evaluate the usability, and discover the potential user interface improvements.

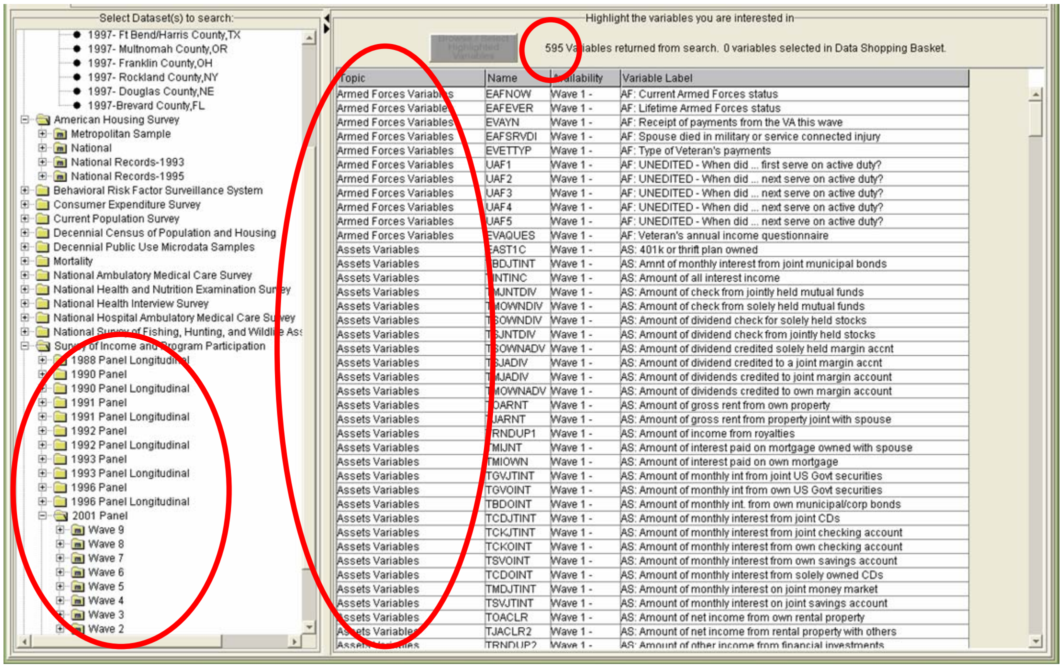
2 PUFs (Public Use Files)

Also known as microdata, public use files (PUFs) are computer-accessible files containing records for a sample of housing units, with information on the characteristics of each housing unit and the people in it. The advantage of releasing microdata instead of precomputed statistics is an increased flexibility and availability of information for users. Microdata files from all Census Bureau demographic surveys (e.g. American Community Survey, American Housing Survey, Consumer Expenditure Survey, Current Population Survey, Long Form of the Decennial Census, Survey of Income and Program Participation, Survey of Program Dynamics, and so on) and other Federal agencies, such as the National Center for Education Statistics, Energy Information Administration, and Internal Revenue Service are today made available online or for a nominal fee to researchers. There are also databases maintained and released by states' departments of motor vehicles (DMVs), health maintenance organizations (HMOs), insurance companies, public offices, commercial organizations, and so on. Most of the PUFs released by Federal agencies have three major characteristics that should be considered in designing user interfaces for a successful disclosure avoidance system.

- **Data Scale:** PUFs are both vertically and horizontally large scale data. PUFs might contain records representing 1 percent or even 5 percent samples of the housing units in the U.S. and the persons in them. Therefore, PUFs often contain more than hundreds of thousands of records. Some PUFs even contain more than one million records if they contain data from repeated interviews. Considering that general display monitors have a few million pixels, it's impossible to project the PUFs records onto a monitor screen all at once without occlusion. In addition to the large number of records, PUFs data files also contain large number of variables. In Figure 1, there are 595 variables (red circle on the right) defined in SIPP microdata. Similarly, most PUFs contain hundreds of variables which are classified by topics such as age, race, sex, marital status, education, industry, occupation, household income, health insurance, and so on. However, the number of variables that can be visualized simultaneously on a display monitor is at most 7 or 8 because there are only a limited number of visual cues that can be used for visualizing multiple variables in 2D or 3D space (e.g. spatial x-y-z location, color, shape, size, and so on).
- **Data Hierarchy:** The variables defined in PUFs usually have hierarchical structures. As shown in Figure 1 (red ellipse in the middle), the variables are classified into several topics, which are often organized by data hierarchy. For example, SIPP has a 5-level data hierarchy structure. Each record in SIPP contains household variables, family and subfamily variables, personal variables from multiple interviews, sample unit

variables which can be further divided into wave variables (each wave data is composed of a few reference month interviews) and reference month variables. There are two different ways to represent the hierarchies in PUFs. PUFs such as SIPP use one big table containing all the variables that come from all data hierarchies. On the other hand, PUFs such as those from ACS (American Community Survey) use separate files so that each file can represent each level of hierarchy. They are linked together through the unique identifiers of upper hierarchy level data. The data hierarchy is often used as a basic unit for expanding or reducing the search space in the disclosure avoidance process. Therefore, the coordinated (synchronized) data viewers for each data hierarchy will facilitate the search and browsing of possible targets.

- **Time-Series Data:** Some PUFs contain data from repeated interviews. This is called time-series data and the data represent a sequence of observations that are ordered in time. There are two kinds of time series data: continuous (an observation at every instance of time such as lie detectors or electrocardiograms) and discrete (an observation at regularly spaced intervals). Most of the PUFs come under the discrete time series category. Figure 1 (red ellipse on the bottom left) shows that SIPP contains multiple years of microdata and each year data is composed of several waves (four-month-interviewing cycles). The change of attribute values over time is often used as a key for finding possible targets. Therefore, appropriate visualization for the change of attribute values over time as well as novel user interfaces for extracting those changes will facilitate the task. In addition to the direct data manipulation, the coordinated data visualizations with time-series visualization can make the disclosure avoidance process much easier.


Select Dataset(s) to search:
Highlight the variables you are interested in
Browse / Select Highlighted Variables

- 1997- Ft Bend/Harris County,TX
- 1997- Multnomah County,OR
- 1997- Franklin County,OH
- 1997- Rockland County,NY
- 1997- Douglas County,NE
- 1997-Brevard County,FL

American Housing Survey
Metropolitan Sample
National
National Records-1993
National Records-1995
Behavioral Risk Factor Surveillance System
Consumer Expenditure Survey
Current Population Survey
Decennial Census of Population and Housing
Decennial Public Use Microdata Samples
Mortality
National Ambulatory Medical Care Survey
National Health and Nutrition Examination Survey
National Health Interview Survey
National Hospital Ambulatory Medical Care Survey
National Survey of Fishing, Hunting, and Wildlife As
Survey of Income and Program Participation
1988 Panel Longitudinal
1990 Panel
1990 Panel Longitudinal
1991 Panel
1991 Panel Longitudinal
1992 Panel
1992 Panel Longitudinal
1993 Panel
1993 Panel Longitudinal
1996 Panel
1996 Panel Longitudinal
2001 Panel
Wave 9
Wave 8
Wave 7
Wave 6
Wave 5
Wave 4
Wave 3
Wave 2

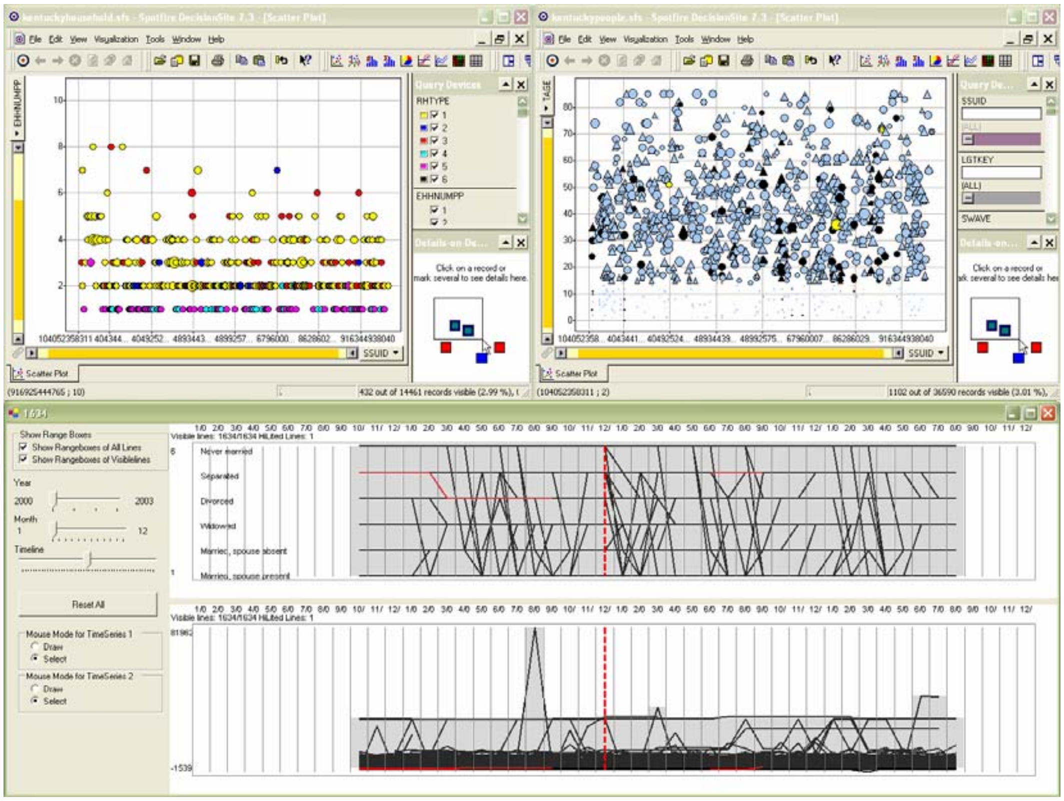
595 Variables returned from search. 0 variables selected in Data Shopping Basket.

Topic	Name	Availability	Variable Label
Armed Forces Variables	EAFNOW	Wave 1 -	AF: Current Armed Forces status
Armed Forces Variables	EAFEVER	Wave 1 -	AF: Lifetime Armed Forces status
Armed Forces Variables	EVAYN	Wave 1 -	AF: Receipt of payments from the VA this wave
Armed Forces Variables	EAFSRVDI	Wave 1 -	AF: Spouse died in military or service connected injury
Armed Forces Variables	EVETTYP	Wave 1 -	AF: Type of Veteran's payments
Armed Forces Variables	UAF1	Wave 1 -	AF: UNEDITED - When did ... first serve on active duty?
Armed Forces Variables	UAF2	Wave 1 -	AF: UNEDITED - When did ... next serve on active duty?
Armed Forces Variables	UAF3	Wave 1 -	AF: UNEDITED - When did ... next serve on active duty?
Armed Forces Variables	UAF4	Wave 1 -	AF: UNEDITED - When did ... next serve on active duty?
Armed Forces Variables	UAF5	Wave 1 -	AF: UNEDITED - When did ... next serve on active duty?
Armed Forces Variables	EVAQUES	Wave 1 -	AF: Veteran's annual income questionnaire
Assets Variables	EAST1C	Wave 1 -	AS: 401k or thrift plan owned
Assets Variables	TBDJTINT	Wave 1 -	AS: Amnt of monthly interest from joint municipal bonds
Assets Variables	TINTINC	Wave 1 -	AS: Amount of all interest income
Assets Variables	TMJNTDIV	Wave 1 -	AS: Amount of check from jointly held mutual funds
Assets Variables	TMOWNDIV	Wave 1 -	AS: Amount of check from solely held mutual funds
Assets Variables	TSOWNDIV	Wave 1 -	AS: Amount of dividend check for solely held stocks
Assets Variables	TSJNTDIV	Wave 1 -	AS: Amount of dividend check from jointly held stocks
Assets Variables	TSOWNADV	Wave 1 -	AS: Amount of dividend credited solely held margin accnt
Assets Variables	TSJADIV	Wave 1 -	AS: Amount of dividend credited to a joint margin accnt
Assets Variables	TMJADIV	Wave 1 -	AS: Amount of dividends credited to joint margin account
Assets Variables	TMOWNADV	Wave 1 -	AS: Amount of dividends credited to own margin account
Assets Variables	TOARNT	Wave 1 -	AS: Amount of gross rent from own property
Assets Variables	TJARNT	Wave 1 -	AS: Amount of gross rent from property joint with spouse
Assets Variables	TRNDUP1	Wave 1 -	AS: Amount of income from royalties
Assets Variables	TMIJNT	Wave 1 -	AS: Amount of interest paid on mortgage owned with spouse
Assets Variables	TMIOWN	Wave 1 -	AS: Amount of interest paid on own mortgage
Assets Variables	TGVJTINT	Wave 1 -	AS: Amount of monthly int from joint US Govt securities
Assets Variables	TGVOINT	Wave 1 -	AS: Amount of monthly int from own US Govt securities
Assets Variables	TBDOINT	Wave 1 -	AS: Amount of monthly int. from own municipal/corp bonds
Assets Variables	TCDJTINT	Wave 1 -	AS: Amount of monthly interest from joint CDs
Assets Variables	TCKJTINT	Wave 1 -	AS: Amount of monthly interest from joint checking account
Assets Variables	TCKOINT	Wave 1 -	AS: Amount of monthly interest from own checking account
Assets Variables	TSVOINT	Wave 1 -	AS: Amount of monthly interest from own savings account
Assets Variables	TCDOINT	Wave 1 -	AS: Amount of monthly interest from solely owned CDs
Assets Variables	TMDJTINT	Wave 1 -	AS: Amount of monthly interest on joint money market
Assets Variables	TSVJTINT	Wave 1 -	AS: Amount of monthly interest on joint savings account
Assets Variables	TOACLR	Wave 1 -	AS: Amount of net income from own rental property
Assets Variables	TJACLR2	Wave 1 -	AS: Amount of net income from rental property with others
Assets Variables	TRNDUP2	Wave 1 -	AS: Amount of other income from financial investments

Figure 1: Data Ferret (a tool for extracting microdata through the internet, http://dataferrett.census.gov) user interface shows the list of 595 variables defined in SIPP microdata, which are classified into several hierarchical topics. The selected variables of microdata can be extracted by timeline such as year and wave.

3 IDFinder

The prototype design of the IDFinder user interface was mainly derived from the data analysis on PUFs and the requirement analysis for disclosure avoidance tasks. Overall visualization design was based on the widely cited principle, usually known as the visual-information-seeking mantra: “Overview First, Zoom and Filter, then Details on Demand” (Card, Mackinlay, & Shneiderman, 1998; Shneiderman & Plaisant 2004). Direct manipulation and dynamic query techniques (Ahlberg, Williamson, & Shneiderman, 1992), (Williamson & Shneiderman, 1992) were also used to support users’ perceptual capabilities in using a data mining tool. Figure 2 shows the overall IDFinder user interface visualizing SIPP microdata from a state in the U.S. In this example, IDFinder consists of two data viewers and a time-series viewer. Each data viewer represents different data hierarchies (the data viewer on the top left corner visualizes household variables while the data viewer on the top right corner visualizes personal variables). Users can add or remove data viewers so that they can focus only on the data hierarchies in which they are interested. The time-series viewer on the bottom visualizes the change of attribute values over time. Just as with the top data viewers, users can add or remove the time-series viewers based on data hierarchies. Each time-series viewer can visualize multiple variables simultaneously if they are in the same data hierarchy. The time-series viewer can also be used for extracting data items whose attribute values are in between specified values during a specified time period. All the viewers in IDFinder are tightly coupled so that any changes such as selection and filtering in one viewer can be reflected on the other viewers.



(916925444765 ; 10)

432 out of 14461 records visible (2.99 %),

(104052358311 ; 2)

1102 out of 36590 records visible (3.01 %),

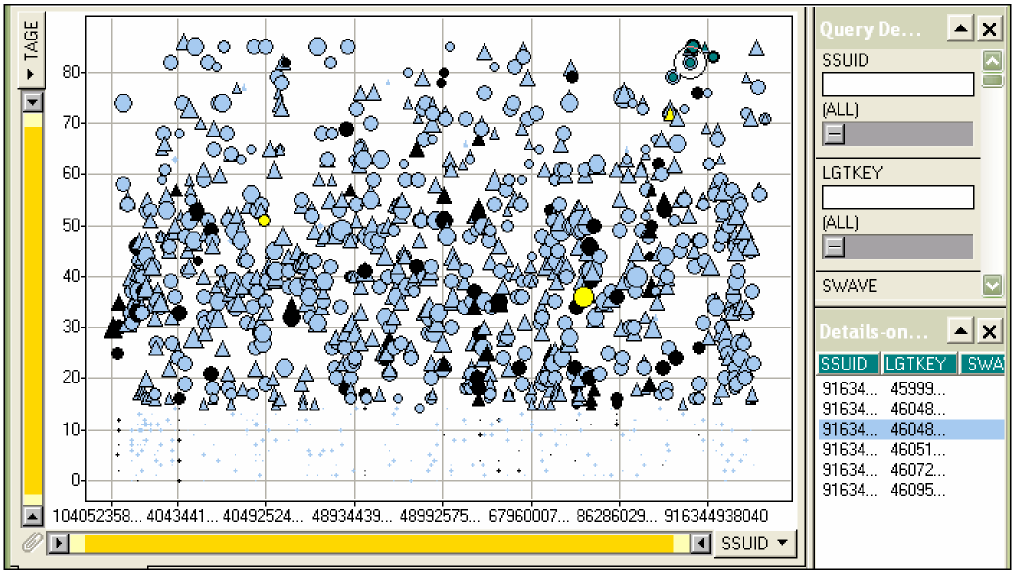
Visible lines: 1634/1634 HiLited Lines: 1

Visible lines: 1634/1634 HiLited Lines: 1

Figure 2: IDFinder user interface with two hierarchical data viewers (household data viewer – top left, and personal data viewer – top right) and a time series viewer on the bottom

3.1 Data Hierarchy-Based Viewer

The SpotFire (http://www.spotfire.com) visualization component has been used to implement the data viewer in IDFinder. Each data viewer represents one data hierarchy in IDFinder. Figure 3 shows an example of a data viewer visualizing the personal variables in SIPP microdata. The data viewer in IDFinder allows users to choose the visual cues for the variables. For example, in Figure 3, the y-axis represents the age of people, the x-axis represents a unique household identification number, so that items located at the same x position means people are living in the same household. Color represents race (light blue:white, black:black, yellow:asian), shape represents sex (triangle:male, circle:female), and the size of a symbol represents the education level (bigger:higher education). The visualized symbols can be filtered dynamically by query slider bars (Ahlberg & Shneiderman, 1994) located at the top right corner of the data viewer, or they can be zoomed in or out by adjusting the slider bars located at the x and y axes. Users can see the detailed information (all the variables) of the selected symbols at the bottom right corner of the data viewer. Data viewers can be added or removed anytime depending on which data hierarchy a user wants to see. With the design of separate data viewers for each data hierarchy, the total number of the variables that can be visualized simultaneously in a display screen can be increased up to the number of data viewers times the number of visual cues.

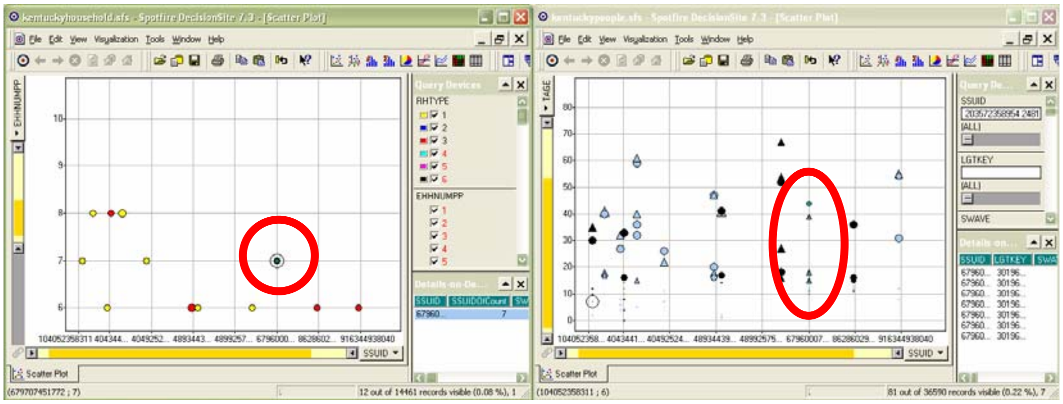


SSUID	LGTKEY	SWA
91634...	45999...	
91634...	46048...	
91634...	46048...	
91634...	46051...	
91634...	46072...	
91634...	46095...	

Figure 3: IDFinder data viewer visualizing personal variables of SIPP microdata. Each symbol (circle or triangle) in the data viewer represents a respondent in the SIPP survey.

3.2 Tightly Coupled Viewer

Multiple coordinated visualizations enable users to rapidly explore complex information (North, & Shneiderman, 2001). In IDFinder, all the data viewers and the time-series viewers are designed to be tightly coupled so that any data changes such as selection and filtering in one viewer can be reflected on the other viewers. Figure 4 shows that users filtered out the households whose size is smaller than 6 from the household data viewer (on the left) and consequently, the personal data viewer (on the right) shows only the people living in the corresponding households. IDFinder also supports brushing. When the mouse is over a household item in household data viewer (red circle on the left), the people in this household are highlighted in the personal data viewer (red ellipse on the right). All of the selection and filtering operations are bi-directional and all of the queries built through dynamic query widgets in each data viewer are joined conjunctively. For example, if the x-axis slider in a personal data viewer is adjusted only to show the people whose age is over 50, then the household data viewer will display only the households whose size is greater than or equal to 6 and that contain a household member whose age is over 50.



kentuckyhousehold.sfs - Spotfire DecisionSite 7.3 - [Scatter Plot]

(679707451772 ; 7)

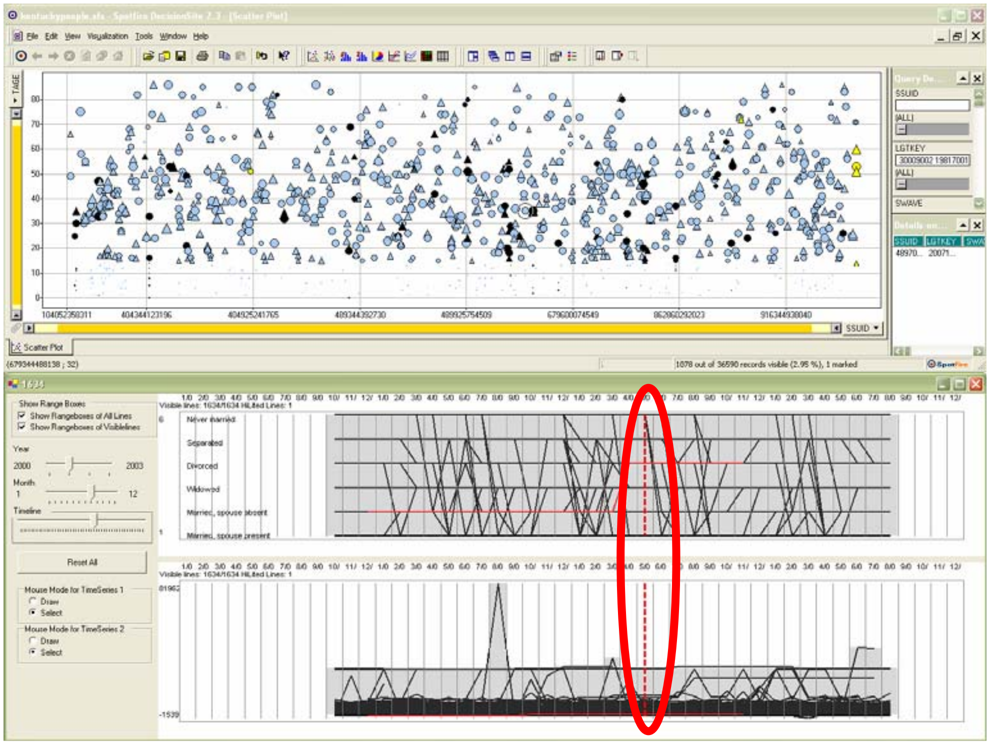
12 out of 14461 records visible (0.08 %), 1

kentuckypeople.sfs - Spotfire DecisionSite 7.3 - [Scatter Plot]

(104052358311 ; 6)

81 out of 36590 records visible (0.22 %), 7

Figure 4: Two tightly coupled data viewers in IDFinder. With this design, users can perform different types of coordination tasks such as brushing and drill down.



kentuckypeople.sfs - Spotfire DecisionSite 7.3 - [Scatter Plot]

(679344488138 ; 32)

1078 out of 36590 records visible (2.95 %), 1 marked

Visible lines: 1634/1634 HiLited Lines: 1

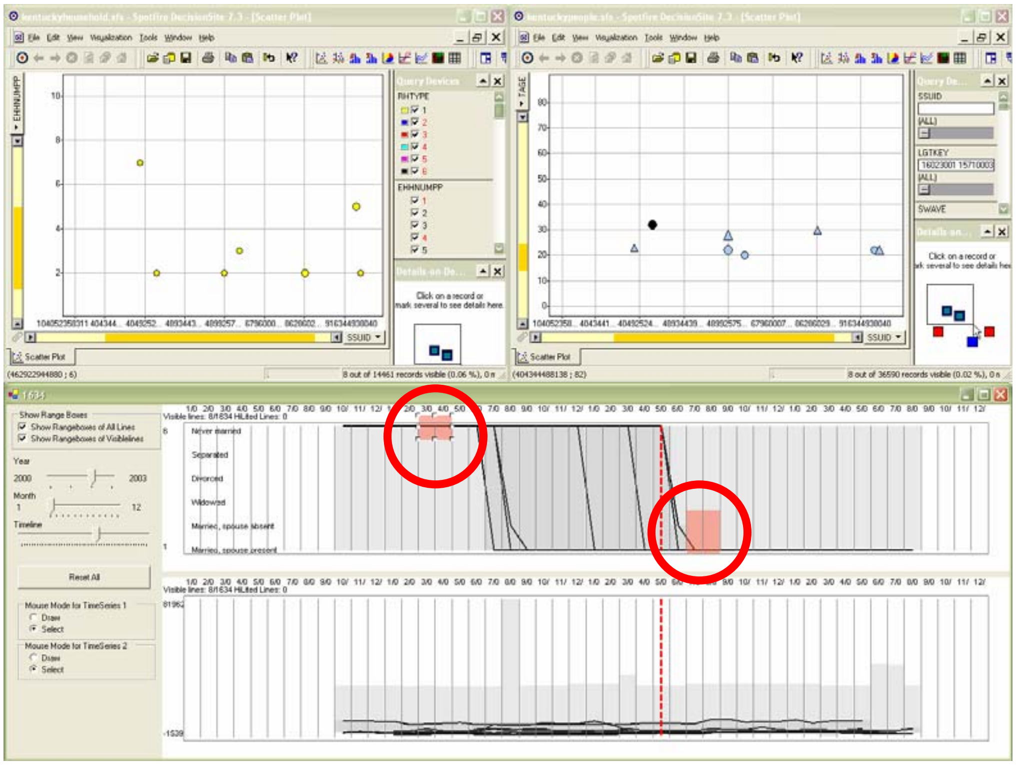
Figure 5: Tightly coupled data viewer and time series viewer. By moving the timeline (red vertical dashed line in the time-series viewer on the bottom) to a specific time in the time series viewer, users can see the attribute values of data items at the specified time in the data viewer.

3.3 Time-Based Visualization

Each data viewer can be combined with a time-series viewer to visualize the changes of variables over time. Users can select multiple variables simultaneously in the time series viewer as long as the selected variables are in the same data hierarchy. In Figure 5, a personal data viewer is combined with a time series viewer that visualizes two personal variables (marital status and monthly income) changing over time. The multivariate time-series visualizations in a time series viewer are also tightly coupled with each other as well as with the other data viewers. When the mouse is over an item in the data viewer, the corresponding item in the multivariate visualizations of the time series viewer is highlighted and vice versa. As shown in Figure 5, there is a timeline (the red vertical dashed line in the time-series viewer) implemented in the time-series viewer. This timeline is designed to be synchronized with the data viewer so that the data viewer can show only the data items at the time that is specified by the timeline. The timeline can be played sequentially for a certain period of time to show the animated change of attribute values in the data viewers.

3.4 Time Series Viewer

The design of the time-series viewer is based on TimeSearcher (Hochheiser & Shneiderman, 2004) which is a prototype tool for interactive querying and exploration of time-series data.



File Edit View Visualization Tools Window Help
Query Devices
RHTYPE
EHHNUMPP
Details on De...
Click on a record or mark several to see details here.
SSUID
Scatter Plot
8 out of 14461 records visible (0.06 %), 0 m
8 out of 36590 records visible (0.02 %), 0 m
TAGE
LGTKEY
SWAVE
Show Range Boxes
Show Rangeboxes of All Lines
Show Rangeboxes of Visiblelines
Year
2000
2003
Month
1
12
Timeline
Reset All
Mouse Mode for TimeSeries 1
Draw
Select
Mouse Mode for TimeSeries 2
Draw
Select
Visible lines: 8/1634 HiLited Lines: 0
Never married
Separated
Divorced
Widowed
Married, spouse absent
Married, spouse present

Figure 6: Time-series viewer (on the bottom) in IDFinder. Users can build interactive queries by drawing timeboxes (red circles) on the time series viewer. The query results are reflected on the other data viewers.

Queries are built using timeboxes: a graphical, direct-manipulation metaphor for the specification of queries over time-series datasets. These timeboxes support interactive formulation and modification of queries, thus speeding the process of exploring time-series data sets and guiding data mining. The design of the time-series viewer has been improved from TimeSearcher by supporting multivariate time-series data and adding a timeline so that it can be synchronized with data viewers. Figure 6 shows an example of how to build a query using the time-series viewer. In this example, users want to find people whose marital status is changed from “never married” to “married” between April 2001 and June 2002. The query results are synchronized with the data viewers so that the personal data viewer can show the people who satisfy this condition and the household viewer can show the households containing those people.

4 Usability Study

Studying the use of IDFinder is important for three reasons: To evaluate the usability and benefits of IDFinder in terms of performance and user satisfaction; to discover potential user interface improvements; and to gain a deeper understanding about disclosure avoidance tasks and IDFinder’s applicability to other tasks. Because of the difficulties in measuring the efficiency of the disclosure avoidance process, qualitative study was performed mainly through subject interviews. Subjects were given the opportunity to observe the use of the system or freely explore the system, describe problems with the IDFinder interface, and offer suggestions for improvement. Overall, the subjects were largely successful in grasping the concepts of multiple data viewers based on data hierarchies and using the IDFinder interface, indicating satisfaction with the direct manipulation and dynamic queries for exploring microdata. They expressed a preference for the hierarchical data visualizations and the timeline based dynamic visualization for the time-series data. On the other hand, some subjects were concerned about the scalability of IDFinder because the sizes of the data sets they have to handle for disclosure avoidance tasks are much larger than that of the test data set. However, they also pointed out that the search space could be narrowed down to the size of the test data set if there is a way to select the part of a microdata file by sub-categorization, for example, by state in the U.S. The study also helped to identify improvements that reduce the need for query specification of time-series and provide a control panel. During the usability study, some subjects figured out that the microdata contained data coding errors (for example, a person’s marital status was changed from “divorced” to “never married”), and mentioned that IDFinder could be used as a data editing tool for detecting the erroneous attribute values from microdata.

5 Conclusions and Future work

This paper introduces a prototype disclosure avoidance system, IDFinder, which allows the disclosure avoidance researchers to rapidly explore the microdata and find any possible targets of re-identifiaction. IDFinder enables users to dynamically explore the microdata based on data hierarchies and the timeline. Tightly coupled data and time-series viewers as well as direct manipulation and dynamic query techniques facilitate the rapid, incremental and reversible search for data. The usability study on the IDFinder interface revealed benefits, scalability issues, and applicability to other tasks. Subjects succeeded in grasping the concept of data visualization based on data hierarchy and timeline, and using the IDFinder interface. They expressed a preference for the direct data manipulation and dynamic query capabilities. Future research includes designing formal metrics to measure the effectiveness of the disclosure avoidance system and the quantitative user studies for evaluating the effectiveness of the current prototype and suggesting design improvements. In conclusion, one of the key contributions of this paper is to present a good case study of how innovative information visualization can help users solve a problem in the real world.

6 References

Ahlberg, C., & Shneiderman B. (1994). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, *Proceedings of the ACM CHI '94 Conference*, 313-317.

Ahlberg, C., Williamson, C., & Shneiderman, B. (1992). Dynamic Queries for Information Exploration: An Implementation and Evaluation. *Proceedings of the ACM CHI '92 Conference*, 619-626.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1998). Reading in Information Visualization: Using Vision to Think. California: Morgan Kaufmann Publishers, Inc.

Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee (2002). Identifiability in Microdata Files, Office of Management and Budget.

Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology. Office of Management and Budget.

Goldstein, J., & Roth, S. F. (1994). Using Aggregation and Dynamic Queries for Exploring Large Data Sets. *Proceedings of the ACM CHI '94 Conference*, 23-29.

Hawala, S., (2003). Microdata Disclosure Protection Research Experiences at the U.S. Census Bureau. *Proceedings of Workshop on Microdata, Sweden*. Also available at http://www.census.gov/srd/sdc/microdataprotection.pdf

Hawala, S., Zayatz, L., & Rowland, S. (2004). American FactFinder: Disclosure Limitation for the Advanced Query System, *Journal of Official Statistics*, 20(1), pp. 115-124

Hochheiser, H., & Shneiderman, B. (2004). Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration. *Information Visualization* 3(1), 1-18.

North, C., & Shneiderman, B. (2001). Component-Based, User-Constructed, Multiple-View Visualization, *Extended Abstracts of CHI '2001*, 201-202.

Rasinski, K. A., & Wright, D. (2000). Practical Aspects of Disclosure Analysis. *Of Significance: A Topical Journal of the Association of Public Data Users*, 2 (1), 35-41.

Ruggles, S. (2000). Foreword – A Data User’s Perspective on Confidentiality. *Of Significance: A Topical Journal of the Association of Public Data Users,* 2 (1), 1-5.

Samarati, P. (2001). Protecting Respondents’ Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, 13 (6), 1010-1027.

Shneiderman, B., & Plaisant, C. (2004). Designing the User Interface (4th ed.). Massachusetts: Addison Wesley.

Williamson, C., & Shneiderman, B. (1992). The Dynamic Home Finder: Evaluating Dynamic Queries in a Real-Estate Information Exploration System. *Proceedings of ACM SIGIR '92 Conference*, 338-346.