Data Science as a Foundation of Integrating and Enriching Administrative Data: The Case of Constructing Taiwan Indigenous Peoples open research Data (TIPD) Based on Taiwan Household Registration Administrative Data

Ji-Ping Lin, Ph.D.
RCHSS, Academia Sinica
Taipei, Taiwan 115
E-mail: jplin@sinica.edu.tw
Office phone: +886-2-27898139

(1st draft, August 2016)

Abstract Embedded in collecting, cleaning, cleansing, processing, & exploring exploding individual digital records is data science. Data science is by no means a new field of science. Rather, it is multidisciplinary in essence and consists of three necessary components: (1) hacking skills, (2) advanced mathematics and statistics knowledge and skills, & (3) domain knowledge expertise. The research is based on a four-year Joint Research Agreement between Academia Sinica and Council of Taiwan Indigenous Peoples. The paper aims to demonstrate how data science, i.e., integration of hacking skills, advanced math./statistics knowledge & skills, and domain knowledge expertise, is applied to construct Taiwan Indigenous Peoples Open Research Data (TIPD, see http://TIPD.sinica.edu.tw, and https://osf.io/e4rvz/) based on Taiwan Household Registration administrative micro data. This paper reports the progress and efforts of Taiwan academicians struggling to construct contemporary TIPs Data (or TIPD) by integrating the micro data of 2000 Taiwan population census and household registration system, using state-of-the-art data science technology like geocoding of primary statistical areas, record linkage to distinguish natural & social increases, using high-performance cluster computers (HPC) to construct micro genealogy to study pattern of inter- & intra-ethnic marriage and level of integration, etc. The research demonstrates the foundation of data science in processing/enriching administration data by an example of constructing interpersonal relationship databank & exploring how formation of an individual ethnic identity can be traced back by comparing child ethnicity with type of parental ethnic marriage practice and parents' ethnicity information. The research potential contribution lies in using methods embedded in the disciplines of data science that overcome the barriers of legal & ethic issues and thus allows to unleash research creativity by using open-sourced administration data.

Keywords: administrative data, data science, multi-disciplinary integration, open research data TIPD

1 Why the Research & Importance of Data Science

In the past five year, the term "big data" has been very popular in academics & industry (National Research Council 2013). Although there are some criticisms on "big data" as hypo, it can't be denied that in the past 15 years, we are experiencing a new phase of data revolution, mainly seen in humanities and social sciences, which is reshaping our societies in unprecedented ways. Contemporary data revolution and widely available massive/big micro data sets could be attributed largely to the emergence of IoTs (internet of things) and personal mobile devices like smart phones & tablets that offer abundant spatiotemporal individual streaming digital traces. Such countless micro digital records enable us to explore the dynamics & causation & interaction of aggregate phenomena in a more precise way and produce more insights for decision making (Schmidt and Cohen 2013).

Embedded in collecting, cleaning, cleansing, processing, & exploring exploding individual digital records is data science (O'Neil and Schutt 2014). Data science is by no means a new field of science. Rather, it is multidisciplinary in essence and consists of three necessary components: (1) hacking skills, (2) advanced mathematics and statistics knowledge and skills, & (3) domain knowledge expertise. It is worthy of stressing that hacking skills are nothing to do with illegal hackers' activities. Hacking skills refer to the skills & ability of manipulating & integrating digital infrastructure that include hardware setting, operating system, programming languages, & software. In short, data science is based on real world domain knowledge expertise and refers to the extraction of knowledge from data, with the main goals being to extract meaning from data and to produce data products. It employs techniques and theories drawn from many fields within the broad areas like mathematics, statistics, and information technology, including signal processing, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high performance computing.

It has been recognized that issues corresponding to administrative data are quite different to those associated with survey data (Herzog, Scheuren, and Winkler 2007). It is also accepted by many academicians that Researchers have benefitted from richer and more reliable sources of administrative data that provide us with deeper insights and much broader vision than survey data about policy and socioeconomic behavior. Conventional administrative data gathered for administrative purposes by government agencies essentially resemble contemporary big/massive micro data gathered by industry for commercial purpose, although data collecting media tend to differ from each other. Because data science is playing a crucial role in contemporary data revolution in industry, it is worthy of noting that little attention has been paid in the sense that methodology of data science might have great potential to tackle administrative data in terms of data collecting, parsing, cleaning, cleansing, validating, privacy-protecting, quality-assessing, reorganizing, processing, exploring, analyzing & enriching issues (Herzog, Scheuren, and Winkler 2007; Wood 2011).

The research is based on a four-year Joint Research Agreement between Academia Sinica and Council of Taiwan Indigenous Peoples. The paper aims to demonstrate how data science, i.e., integration of hacking skills, advanced math./statistics knowledge & skills, and domain knowledge expertise, is applied to construct Taiwan Indigenous Peoples Open Research Data (TIPD, see https:// https://osf.io/e4rvz/) based on Taiwan Household Registration administrative micro data. This paper reports the progress and efforts of Taiwan academicians struggling to construct contemporary TIPs Data (or TIPD) by integrating the micro data of 2000 Taiwan population census and household registration system, using state-of-the-art data science technology like geocoding of primary statistical areas, record linkage to distinguish natural & social increases, using high-performance cluster computers (HPC) to construct micro genealogy to study pattern of inter- & intra-ethnic marriage and level of integration, etc.

This paper demonstrates the foundation of data science in processing/enriching administration data by an example of constructing interpersonal relationship databank &

exploring how formation of an individual ethnic identity can be traced back by comparing child ethnicity with type of parental ethnic marriage practice and parents' ethnicity information. The paper is organized as follows. Section 2 provides a brief literature review on ethnic identity and marriage practice and their relationships. Section 3 is the backgrounds about TIPs. Section 4 introduces (1) source data sets and methodological foundation in constructing micro genealogy for research, and (2) definition of types of ethnic marriage practice and sources of ethnic identity formation, and variables used to associate research theme. It also briefly introduces skills of data manipulation and integration, statistical computing, and computer programming used to synthesize research data. Section 5 provides research results, including descriptive analyses, patterns and associates of ethnic marriage practice, and sources of individual ethnic identity formation. In the end, Section 6 concludes the research and offers some discussions.

2 A Literature Review on Research Theme

The research is theoretically grounded by the following literature. Ethnic identity is closely linked to various types of social integration and cohesion that have long been accepted as crucial determinants for social and political stability (e.g. Chan & To and Chan 2006; Koopmans and Schaeffer 2016; Laurence 2011; Nagel 1994; Rochelle 2015; Smits 2010). A number of scholars in various disciplines of social sciences have developed theoretical and conceptual frameworks of ethnic identity in which a broad array of definitions have been carefully proposed for and a number of methodological options are devised rigorously to measure identity in the context of components, dynamics, structure and formation of ethnic identity (Erikson 1968; Hogg, Terry, and White 1995; Phinney and Ong 2007; Stryker and Burke 2000; Tajfel 1981; Schwartz et al. 2009; Turner et al. 1987).

There is no unique definition of ethnic identity, but the definition by Tajfel (1981, p. 255) that ethnic identity is an individual self-concept deriving from personal knowledge, value, and emotional attachment of membership in a social group clearly specifies the multidimensionality features of ethnic identity. Phinney (1990) concludes that such situation leads to measure of ethnic identity being widely discrepant and thus empirical results being very inconclusive and difficult to make comparisons across studies, leading to the need for efforts in order to offer a measurement method embedded in theoretical ground to reflect components of ethnic identity. According to Phinney (1990) and Phinney and Ong (2007), the components of ethnic identity consists of self-identification, self-categorization and labeling, exploration, ingroup attitudes, values and beliefs, commitment and attachment, and ethnic behaviors.

In similar situation, Abdelal et al. (2009) point out that, based on their survey of literature review on identity, the most widely used methods for measuring identity in general include surveys, content analysis, discourse and ethnography, cognitive mapping, and experiments. Nevertheless, they find out that "…[w]e did not discover any systematic links between these methods and the types of identity they were used to measure, although nearly all studies of identity included some sort of case study." (Abdelal et al. 2009 pp. 4). In terms of methodology, they argue that "[i]dentity scholarship has so far limited itself to a somewhat narrow methodological band, taking little notice of newer, less traditional options. We are proponents of methodological eclecticism, particularly with regard to identity work." (p.9, para. 2)

Scholars are aware that ethnic identity is not static, but very dynamic in the sense that individual ethnic identity tends to change over time and context within an individual's life course. The formation of an individual ethnic identity serves as an outcome of ethnic identity development that refers to the process of development of an individual ethnic identity from an unexamined to an achieved status through a period of exploration (Erikson 1968; Phinney 1990; Schwartz 2001). It is worthy of stressing that a number of researches have indicated that the development of ethnic identity is closely related to the feelings of self-esteem, particularly for the adolescents (Phinney and Chavira 1992; Smith et al. 1999; Verkuyten and Thijs 2004).

Factors affecting ethnic identity formation consist of parental ethnic socialization, family socioeconomic status (SES), community context. In a similar way, Chandra (2009) points out that from the perspective of constructivism, each individual has a choice set of multiple ethnic identities, of which one ethnic identity can be activated in any given context and the activated ethnic identity can change over time due to institutional changes in political and economic outcomes. One important implication for data collection is that a distinction should be made between ethnic structure and ethnic practice, the former referring to the set of potential ethnic identities that characterize a population and the later to the sec of identities that are actually activated by the population.

Scholars are aware that ethnic attitudes and behaviors of parents are crucial in shaping child ethnic identity, because parents will transmit information, values, and perspectives about ethnicity to their children. Such practices are termed as parents' ethnic socialization and it has been confirmed that parents' ethnic socialization practices have profound effect in shaping children's ethnic identity formation (Hughes 2003; Hughes et al. 2006; Spencer and Markstrom-Adams 1990). Hughes et al. (2006) indicate that "parental practices of ethnic socialization are shaped by individual and group characteristics and by characteristics of the contexts in which parents and children operate." (Hughes et al. 2006, p.757). Based on literature review, they further specify five demographic and two contextual factors that have been widely investigated. The five demographic factors include (1) age and gender of children that serve as proxies for the capacity to understand parental messages, and (2) socioeconomic status and immigration status and identity of parents that serve as proxies for parents' vision in the world, while contextual factors include region and neighborhood (or community) that serve as proxies for ethnic composition and interethnic relations that may shape parents' ethnic socialization messages and discrimination experiences of parents and children.

One important dimension embedded in parental ethnic socialization is the practice of parental ethnic marriage. A rich body of researches on ethnic marriage practice suggests that intermarriage is a good measure reflecting ethnic relationships and social structure; intermarriage is also regarded as an effective measure for social integration (Blau and Schwartz 1984; Gordon 1964; Kalmijn 1998). Moreover, volume and proportion of intermarriages in a society is positively linked to reducing the likelihood of ethnic conflicts, intermarriage thus serves as an important measure of social cohesion (McDoom and Gisselquist 2015; Monden and Smits 2005). For example, the study by Smits (2010) on intermarriage in former Yugoslavia concludes that intermarriage enhances social cohesion in the sense that more intermarriages was associated with less ethnic violent conflicts. Many studies further suggest that in addition to individual preferences, intermarriage is also subject to the constraint of structural factors (Kalmijn 1998; Monden and Smits 2005).

A rich body of researches has indicated factors that may affect the practice of ethnic marriage. The most noteworthy structural factors that affect intermarriage are ethnic group size and the spatial organization such as the levels of geographic and ethnic segregation and concentration. In terms of ethnic group size effect, smaller groups have higher rates of intermarriage than larger groups, other things equal. Spatial organization reflects disparities of spatial distribution in opportunity for personal contact and thus type of ethnic marriage. It is also found that persons with mixed ethnic backgrounds are less likely to marry persons of the same or similar backgrounds than those with unmixed ethnic backgrounds. Birth cohort, religion, and migration experience also serve as major factors for ethnic marriage practice (Alba and Golden 1986; Kalmijn 1998; Monden and Smits 2005).

In light of the role of family and parents in influencing the development of child ethnic identity formation, it is worthy of paying more attentions to the formation of ethnic identity for those with mixed ethnic backgrounds. From the perspective of rational choice theory, an individual with mixed ethnic backgrounds is associated with a choice set of multiple ethnic identity alternatives within which she/he chooses one ethnic alternative with maximum perceived utility from the ethnic identity choice set. In this regard, collecting data on the ethnicity of parents is a necessary condition for research in dealing with issues of formation of individual ethnic identity, because it allows us to categorize practice of parental ethnic marriage and to distinguish child ethnic identity from parents' ethnicity. Unfortunately, researches in this aspect are relatively less documented, mainly because of the lack of adequate data and difficulties in measuring ethnic identity for people with mixed ethnic backgrounds.

As aforementioned, the practice of parental ethnic marriage serves as a determinant in internalizing an individual's ethnic identity formation. Following this logic, it suggests that if an individual's parental marriage is a practice of the same ethnicity, then the individual will claim for sure her/his parents' ethnicity; on the other hand, if the individual's parental marriage is a practice of mixed ethnicity, the individual's ethnic identity may be claimed as either mother's ethnicity (termed as "matrilineal ethnic identity" thereafter) or father's ethnicity (termed as "patrilineal ethnic identity" thereafter). Moreover, individuals with mixed ethnic backgrounds may not claim her/his ethnicity affirmation before reaching the achieved ethnic identity stage. Consequently, further efforts to examine the relationships between parental ethnic marriage practice and child ethnic identity are expected to help reveal the complex development of individual ethnic identity for persons with mixed ethnic backgrounds.

3 Background as Domain Knowledge in Data Science

The population in study is Taiwan Indigenous Peoples (TIPs). TIPs refer to the Austronesian peoples of Taiwan, whose ancestry can be traced back to approximately 6,000 years ago (Blust 1985; Li 2011:157). The Austronesian-speaking peoples include: (1) Taiwanese aborigines; (2) the majority ethnic groups of East Timor, Indonesia, Malaysia, the Philippines, Brunei, Madagascar, Micronesia, and Polynesia; (3) the Polynesian peoples of New Zealand and Hawaii; (4) the non-Papuan people of Melanesia; and (5) a small group of Austronesians in Singapore, the Pattani region of Thailand, and the Cham area of Vietnam, Cambodia, and Hainan. Most linguists and archaeologists favor the hypothesis that Taiwan is the homeland of the Austronesian-speaking population (Bellwood 1991; Blust 1985; Russell 2009; Shutler and Marck 1975).

There was a rich body of ethnographic, anthropological, archaeological, linguistic, official statistics derived from survey, registration, census records on TIPs before 1940. Unfortunately, the period of 1940-2000 marks as data "Dark Ages" for TIPs due to 1941-45 Pacific War, 1946-1990 political authoritarian rule in fears of communism and communists infiltration. Persistent lack of TIPs data led TIPs to become isolated, marginalized and thus underdeveloped. Taiwan resumed TIPs population census in 2000 and began recording TIPs individual records in household registration system since 2003.

Consequently, contemporary TIPs information and statistics are predominantly compiled from household registration data. Current population of TIPs amounts to around 54 thousand persons which makes up about 2.3% of the whole population of Taiwan, while the Han Chinese serves as the majority ethnic group.

The development toward diversity in various aspects also accelerates migrations of TIPs, mostly rural-to-urban migration, that in turn have profound impact on ethnic identity and marriage practice of TIPs. Based on the author's previous studies with peers on the population and internal migration of TIPs, TIPs are characterized by four features in terms of population distribution and migration: (1) geographically segregated population distribution, (2) being very migratory, with migration being mainly rural-to-urban type, (3) periphery of metropolitan areas serving as main destination choice for TIPs rural-to-urban migrants; (4) weak ability of TIPs migrants to make onward migration and return migration as the main type of migration once repeat migration occurs.

4 Data and Methods

The conceptual framework and operational definitions for ethnic marriage practice and formation of individual ethnic identity are straightforward and easy to understand, but it is a challenge to find ways that enables us to construct real-world measure measurement. In the research, we at first construct a genealogy databank that serves as the first step to solve the measurement issues. The constructed genealogy databank not only contains individual information, but also the information on the individual's parents and spouse. Because the data for research are the whole population of TIPs, the author thus does not use any statistical models with random mechanism in the research.

4.1 The Administration Data

The main research data set is a micro genealogy databank (termed as matched databank thereafter) that is constructed by matching all individual records of the November 2013 TIPs household data (termed as master databank thereafter) with a cumulated TIPs databank that collects TIPs information since 2007 (termed as reference databank thereafter). The master databank is all individual records of TIPs in November 2013 and amounts to 532,617 persons. The reference databank that amounts to 3,153,023 records preserves the original states of TIPs information, including those who have passed away or emigrated out of Taiwan.

Both master databank and reference databank share the same variables of record, because they all originate from Taiwan household data. Variables of record in both databanks include: personal identification number, family name, given name, household ID, household full address (including geographic information on zip code, region name, county name, township name, chun-li name), household name, relationship with household head, parents' names, spouse name (if married), gender, date of birth, education, marital status, birth place, and ethnicity.

It is important to stress that if information on death and emigration is not preserved and maintained in the cumulated databank, we would not be able to go back to the past to construct effective information on family lineage. The cumulated databank thus serves as the main reference source data that enables us to enrich the TIPs information of master databank by means of record linkage through the links of parents' names and spouse name. In short, the incorporation of individual records of parents and spouse from the reference databank into the master databank serves as the foundation of the research.

4.2 In-memory Computing Methods as Key to Construction of Genealogy Databank:

Record matching between master databank and reference databank through the link of parents' names and spouse name involves a lot of computing. To reduce unnecessary searches in the reference databank, the reference databank is sorted by the order of gender, family name, and given name. Furthermore, the sorted reference databank is indexed by a file that records information on the first row of record for each sequence of gender, family name, and given name in the sorted reference databank.

The record linkage between the master databank and the sorted reference databank is implemented by the following procedures: first, for any given individual record in the master databank, use information stored in the index file to acquire the first row information for the name to be matched in the sorted reference databank; second, use information retrieved from the index file to locate the first record in the sorted reference databank; third, search the name to be matched in the sorted reference databank; if the name to be matched is not unique in the sorted reference databank, we choose the record with the maximum likelihood as the person to be matched using information on age and ethnicity; fourth, pick up the matched individual record in the sorted reference databank and insert it at the end of the individual for matching in the master databank.

Based on the above mentioned record searching and matching procedures, the research matches each individual in the master databank with the reference databank, with respect to the individual father's name, mother's name, and spouse name. Since the average number of searches in the sorted reference databank for each record matching by name is about 50 thousand, the total number of searches involved in matching records of parents and spouse amounts to around 81 billion times. To accelerate the construction of micro genealogy databank, the research takes advantage of in-memory high performance computing (HPC) techniques. In-memory high performance computing comprises three kernel skills of manipulating digital infrastructure: (1) overclocking CPUs, (2) overclocking internal memory speed, and (3) accelerating I/O bus bandwidth that links CPUs and internal memory.

The research adopts a high performance workstation that has two high-end Xeon 2680 v2 CPUs and 256 GB DDR3-16000 ECC DRAM and a BIOS which allows us to adjust I/O bus and internal memory information transferring speed. In order to control and take full advantage of digital hardware settings that enables us to save substantial amount of computing time, the author develops computing codes in object Pascal programming language and has the codes compiled by Embarcadero RAD Studio XE6 compiler. The programming codes are designed for in-memory computing purpose in the sense all computing tasks of constructing genealogy databank are implemented in computer's internal memory, with CPUs and internal memory being overclocked and I/O bus between CPUs and memory being accelerated.

In short, the processes of constructing genealogy databank are demonstrated in Figure 1. As illustrated in Figure 1, the reference databank is loaded into computer's internal memory and sorted by gender, family name, and given name. The second stage is to analyze the sorted reference databank and build an index file based on gender, family name, and given name. The third stage loads the master databank into memory, and reads individual records sequentially. The fourth stage, using index information, looks for the record in the reference databank with father's, mother's, and spouse's names identical to the father's, mother's, and spouse's names of the individual under process in the master databank. The fifth stage is to append the matched father, mother, and spouse information retrieved from the reference databank to the master databank. When all abovementioned procedures are all finished, the construction of genealogy databank is done.

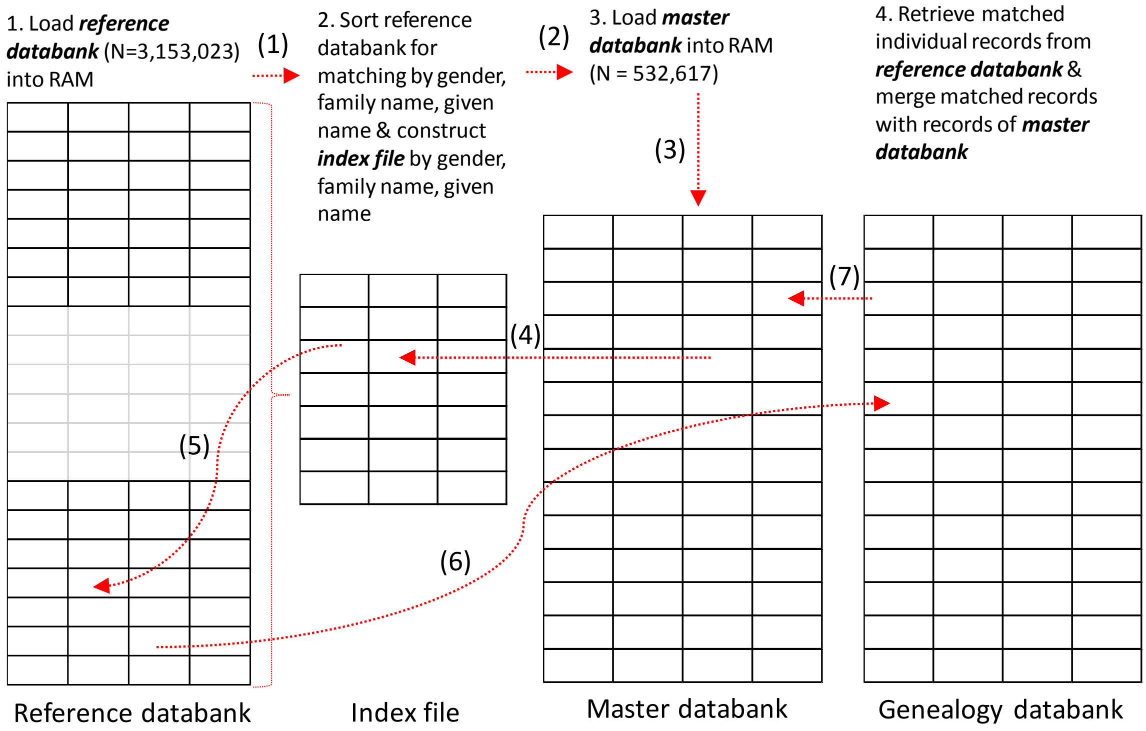

1. Load ***reference databank*** (N=3,153,023) into RAM
(1)
2. Sort reference databank for matching by gender, family name, given name & construct ***index file*** by gender, family name, given name
(2)
3. Load ***master databank*** into RAM (N = 532,617)
(3)
4. Retrieve matched individual records from ***reference databank*** & merge matched records with records of ***master databank***
(7)
(4)
(5)
(6)
Reference databank
Index file
Master databank
Genealogy databank

Figure 1 procedures of record matching using in-memory computing

It is worthy of noting that privacy and research ethics are the first concern of the research. The research data can only be analyzed in a computing environment without any physical communication devices connected to the workstation for computing in the funding agency. However, the research is open to offer computing source programs coded by the author to any interested researchers upon request. Open source of computing codes not only allows other researchers to examine potential coding flaws, but also enables others to improve coding efficiency and promotes application to other similar research.

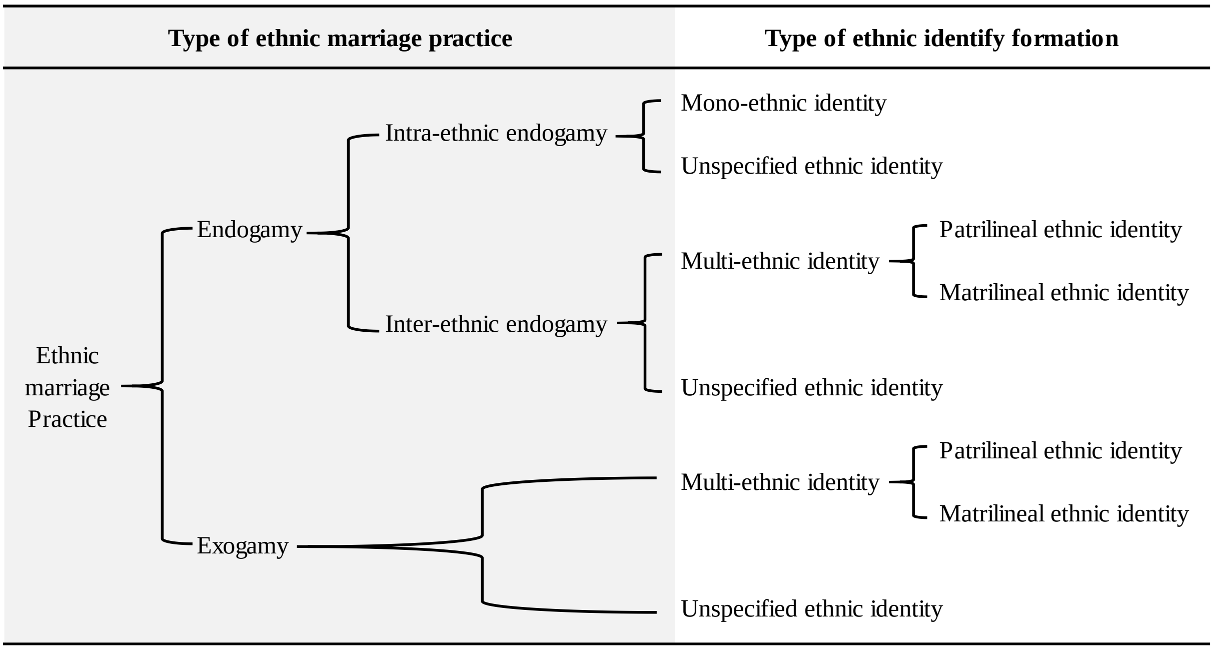
4.3 Definition and Measurement

The research does not use the terms in-marriage and intermarriage to define ethnic marriage type, because both terms won't allow us to distinguish (1) marriage of TIPs persons with partners who are not TIPs and (2) marriage of persons belonging to one specific TIPs ethnicity with persons to other TIPs ethnicity. In the research, instead, type of marriage practice is dichotomized into two categories: endogamy and exogamy. In the paper, a TIPs individual's marriage practice is defined as endogamy if she or he marries the one who belongs to TIPs, otherwise defined as exogamy. In other words, endogamy refers to the marriage practice among the TIPs, while exogamy is the marriage practice between TIPs and those who are not TIPs. The practice of endogamy for a given TIPs couple is further dichotomized into two categories: intra-ethnic endogamy and inter-ethnic endogamy. Intra-ethnic endogamy refers to the endogamy of a TIPs couple with the same TIPs ethnicity. On the other hands, inter-ethnic endogamy refers to the endogamy of a TIPs couple with different TIPs ethnicity.

The main research data allow us to capture individual ethnic information. Ethnic identity of TIPs refers to the individual ethnicity information recorded in the research data.

Ethnicity information in the research data includes the following categories of ethnicity for TIPs population: Amis, Atayal, Paiwan, Bunun, Rukai, Puyuma, Tsou, Saisiyat, Dao, Thao, Kavalan, Truku, Sakizaya, Sediq. Nevertheless, it is worthy of highlighting that if a TIPs individual does not claim her/his ethnic information, the ethnicity information in the research data is recorded as “unspecified ethnicity”.

Table 1 Marriage practice and category of ethnic identity formation


Type of ethnic marriage practice
Type of ethnic identify formation
Ethnic marriage Practice
Endogamy
Intra-ethnic endogamy
Mono-ethnic identity
Unspecified ethnic identity
Inter-ethnic endogamy
Multi-ethnic identity
Patrilineal ethnic identity
Matrilineal ethnic identity
Unspecified ethnic identity
Exogamy
Multi-ethnic identity
Patrilineal ethnic identity
Matrilineal ethnic identity
Unspecified ethnic identity

The research defines four types of individual ethnic formation based on the comparison between individual ethnicity and type of parental ethnic marriage practice. They include (1) mono-ethnic identity, (2) patrilineal ethnic identity, (3) matrilineal ethnic identity, and (4) undetermined ethnic identity. A TIPs individual is defined as having mono-ethnic identity formation if her/his parental marriage practice is intra-ethnic endogamy and her/his ethnicity is identical to her/his parental ethnicity. A TIPs individual is defined as being associated with patrilineal- and matrilineal-ethnic identity formation if her/his (1) parental marriage practice is characterized with exogamy or with inter-ethnic endogamy and (2) individual ethnicity is identical to father’s ethnicity and mother’s ethnicity, respectively. If none of the aforementioned conditions hold, a TIPs individual ethnic identity formation is defined as “undetermined” although information regarding her/his father’s and/or mother’s ethnicity might be available in the constructed genealogy databank. In short, the research uses Table 1 to summarize how the research uses (1) types of parental marriage practice and (2) parental ethnicity and child ethnicity information to define the four types of individual ethnic formation.

5 Results

5.1 Open Data as an Effective Way to Overcome Legal & Ethic Issues

Major outputs of TIPD which are open to the public amount to 7,300 files in number and around 32 GB in size (see https://osf.io/e4rvz/). TIPD are bilingually documented and its content, context, and volume are growing steadily. TIPD now consist of three categories of open research data: (1) categorical data, (2) household structure and characteristics data, and (3) population dynamics data.

Categorical data include two broad dimensions. The first one is contingency tables which are available in PDF, HTML, RTF, XLS formats, while the other is multi-dimensional data which are offered in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats.Household structure and characteristics data consist of three broad dimensions of information: (1) household head information, (2) household member composition information, and (3) household geographical information. They are also available in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats.

Population dynamics data consists of three categories: (1) increased population within a given period of time. It can be further dichotomized into two branches of data, population increase due to birth and due to immigration; (2) decreased population within a given period of time. It can also be divided into two branches of data, population decrease due to death and due to emigration; (3) intact population within a given period of time. It can be distinguished into two categories of population: those who make internal migration and those remaining staying-put. For intact population who make internal migration, internal migration processes such as in-, out-, net, gross migrations are analyzable. Every types of population dynamics data are available in CSV, Excel, dBase, Access, Matlab, Gauss, HTML, JMP, SAS, SPSS, Stata, & Access formats.

The potential applications for research on TIPs based on TIPD include studies on birth, death, migration, residential mobility, life table, marriage, ageing, education, medical care, labor, family, community, etc. TIPD could be used as background data for survey study, including population analysis, sampling design and sampling planning. In short, not only does construction of TIPD which are derived from confidential micro household administrative micro data overcome legal & ethic issues that allow to unleash social creativity in research on TIPs, but it also contributes to shed lights on contemporary Taiwan Indigenous Peoples and human dynamics which have been "invisible" to the world for seven decades.

5.2 Data Visualization and Exploratory Statistics as the First Step Toward Data Science: Characteristics of TIPs

Based on the master databank, this section provides results of descriptive analyses that offer characteristics and insights about TIPs. The research uses Table 2 and Figure 2 to demonstrate the population characteristics and spatial distribution patterns of TIPs. Table 2 provides descriptive statistics regarding population size, share of population residing in metropolitan areas, gender, age, education, and marital status with respect to each ethnic group of TIPs. As shown in Table 2, the four largest ethnic groups (Amis, Atayal, Paiwan, and Bunun) make up 81% of total TIPs population, with ethnic Amis's share of total TIPs being as high as 37%. Except for ethnic Amis that populate mostly either in eastern Taiwan or metropolitan areas of northern Taiwan, most ethnic groups of TIPs are associated with lower share of population residing in metropolitan areas.

Sex ratio (the ratio of male to female population) demonstrates that except for two small ethnic groups (Sediq and Sakizaya), TIPs are characterized by a feature that females outnumber males in terms of total population size. Such feature in sex ratio is particularly noteworthy for ethnic Atayal. This feature in sex ratio is predominantly shaped by two factors: first, male TIPs are associated with a much lower level of life expectancy than national average level of male life expectancy; second, the life expectancy of female TIPs does not differ too much from the national average level of female life expectancy. In terms of educational composition that is measured by share of population with at-least some college, ethnic Thao and Sakizaya have highest share of population with higher education (20.5% and 18.4%, respectively), and it is not surprising to find that ethnic Dao which mainly populate in a small island has a share of only 9.3%. The age structure of TIPs is demonstrated by statistics on share of those aged 14 and less and share of those aged 65 and over in Table 2, with an aim to illustrate population of dependency. The age structure of TIPs is characterized by relatively low share of elderly population (aged 65 and over) and higher share of young population (aged 14 and less), with ethnic Sakizaya as an exception. Table 2 also summarizes statistics on marital status, indicating that the average share of married people is 33.3%.

Table 2 Descriptive statistics on Taiwan indigenous peoples (TIPs) by ethnicity

Ethnicity	Population	Residence	Sex	Education	Age			Marital status			
	Persons (% of total)	% residing in metro areas	sex ratio (males/females*100)	% at-least some college	% aged 14 and less	% aged 65 and over	dependency ratio	% Single	% Married/spoused	% Divorced	% Widowed
Overall	532,617 (100.0)	36.8	95.5	12.5	21.4	6.2	38.1	53.0	33.3	8.9	4.8
Ethnicity											
Amis	197,235 (37.0)	47.7	97.1	11.8	20.2	7.4	38.2	52.3	33.5	9.9	4.3
Atayal	84,551 (15.9)	37.4	90.3	12.6	24.6	5.1	42.1	53.0	33.6	9.1	4.3
Paiwan	94,773 (17.8)	29.5	94.3	12.9	21.4	6.3	38.1	52.3	33.8	8.4	5.5
Bunun	55,165 (10.4)	27.7	94.1	12.4	24.0	3.6	38.0	54.3	34.2	6.3	5.1
Rukai	12,664 (2.4)	26.8	95.2	13.8	19.1	7.8	36.7	51.6	34.8	7.6	6.0
Puyuma	13,075 (2.5)	31.9	94.7	14.4	20.1	6.8	36.8	54.4	30.3	10.2	5.1
Tsou	7,070 (1.3)	23.7	93.2	14.4	21.2	6.7	38.7	50.7	36.6	7.1	5.6
Saisiyat	6,311 (1.2)	42.6	94.5	12.9	23.4	5.4	40.5	55.3	31.7	7.7	5.3
Dao	4,361 (0.8)	6.3	98.6	9.3	20.9	7.0	38.7	54.0	35.7	5.8	4.5
Thao	743 (0.1)	34.2	93.5	20.5	20.7	4.6	33.9	57.3	33.0	6.6	3.1
Kavalan	1,338 (0.3)	39.2	98.8	13.8	18.5	7.8	35.7	54.6	29.7	10.1	5.6
Truku	28,990 (5.4)	19.8	93.6	12.8	24.4	4.7	41.0	55.5	30.3	9.3	4.8
Sakizaya	751 (0.1)	20.8	102.4	18.4	15.4	18.8	52.0	44.3	37.9	10.1	7.6
Sediq	8,626 (1.6)	15.3	101.7	13.6	24.2	6.3	43.9	51.8	35.5	7.6	5.1
Others	16,964 (3.2)	40.4	118.8	14.9	7.8	6.4	16.5	55.4	28.8	11.0	4.7

Note data source: the 2013 year end of Taiwan indigenous peoples household registration data (*N* = 532,617 persons); "Others" of ethnicity include ethnic Saarua, ethnic Kanakanavu, & those who do not specify their ethnicity.

Because existing literature stresses that spatial distribution of ethnic groups is crucial to explain ethnic in-group organization and ethnic inter-group interaction, the research maps the spatial distribution of TIPs residential place with respect to each ethnic group based on the master databank, as shown in Figure 2. The base maps in Figure 2 include two categories: (1) the base map that distinguishes central mountainous areas with non-mountainous areas; and (2) the base map that distinguishes metropolitan areas and non-metropolitan areas. The former is used to demonstrate the effects of physical environments, while the later to illuminate the effects of spatial socioeconomic organization. The metropolitan areas in Figure 2 include the following metropolitans: Taipei, Taoyuan, Taichung, Tainan, Kaohsiung, Hsinchu, and Chiayi, which are categorized on the basis similar to the definition of standard metropolitan statistical area (SMSA).[1]

In Figure 2, Figure 2.a demonstrates that the overall pattern of TIPs population spatial distribution is characterized by three features: (1) rural TIPs population mainly populate in eastern Taiwan and in central mountainous highlands, (2) urban TIPs population are seen to concentrate mostly in northern Taipei and Taoyuan metropolitan areas and partly in southern Kaohsiung and central Taichung metropolitan areas, and (3) TIPs population barely reside in the rural areas of western Taiwan where are mainly populated by ethnic Han and other ethnic

[1] For details about SMSA definition, see https://www.census.gov/population/metro/

groups since the 17th century. The spatial distribution of TIPs by ethnicity is illustrated by the remaining sub-figures of Figure 2. Large ethnic groups exhibit two patterns of spatial distribution: concentrated and scattered type. Small ethnic groups are all characterized by concentrated pattern. Ethnic spatial distribution of TIPs population reveals not only information of residential place, but also embedded social structure and network that indirectly shape ethnic spatial distribution. For example, Figures 2.b, 2.d, and 2.f exhibit spatial pattern of ethnic Amis, Paiwan, and Rukai that are characterized as being highly concentrated, whereas Figures 2.c, 2.e, and 2.j demonstrate that ethnic Atayal, Bunun, and Truku are seen to be characterized by scattered pattern of spatial distribution.

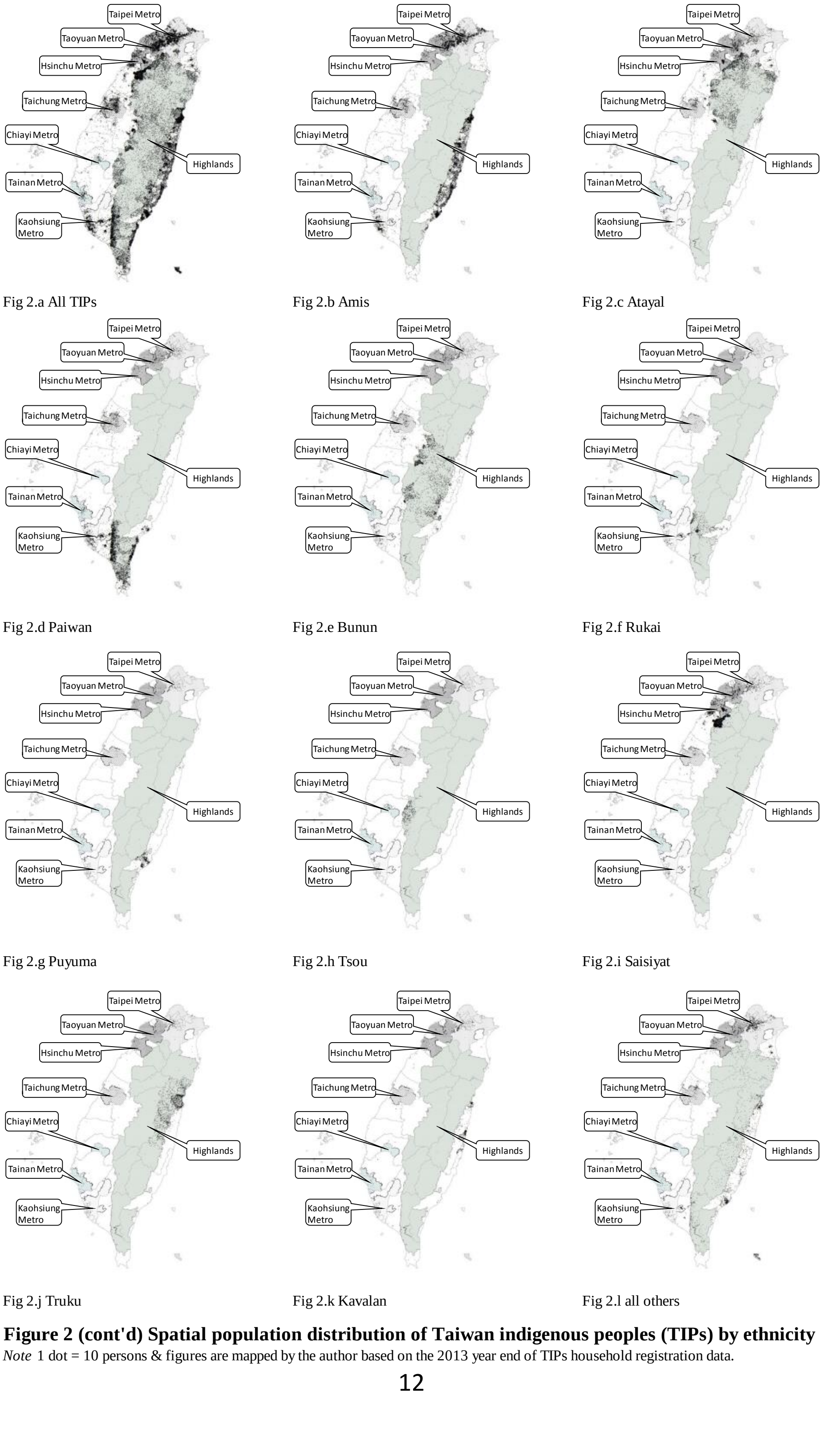


Fig 2.a All TIPs

Fig 2.b Amis

Fig 2.c Atayal

Fig 2.d Paiwan

Fig 2.e Bunun

Fig 2.f Rukai

Fig 2.g Puyuma

Fig 2.h Tsou

Fig 2.i Saisiyat

Fig 2.j Truku

Fig 2.k Kavalan

Fig 2.l all others

Figure 2 (cont'd) Spatial population distribution of Taiwan indigenous peoples (TIPs) by ethnicity

Note 1 dot = 10 persons & figures are mapped by the author based on the 2013 year end of TIPs household registration data.

5.3 Enriching Administration Data Based on HPC and In-memory Computing Methods to Unveil the Information Never Seen before: Pattern of Ethnic Marriage

This section provides findings on (1) patterns of exogamy and endogamy, (2) patterns of intra-ethnic endogamy and inter-ethnic endogamy, and (3) association of important correlates with the aforementioned ethnic marriage practices. Based on previous findings of researches, the correlates used in the research include individual ethnicity, type of parental ethnic marriage practice, gender, age, education, and type of residential place.

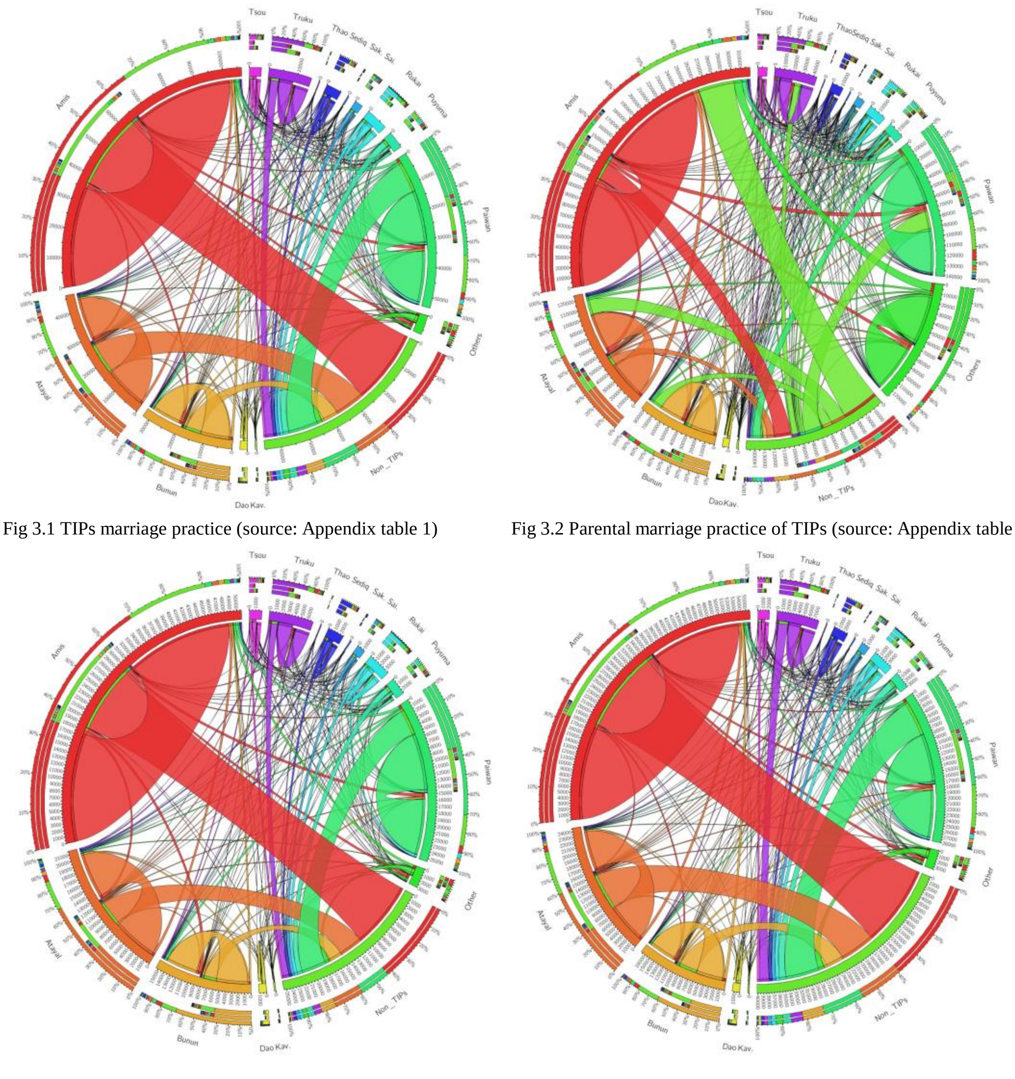


Fig 3.1 TIPs marriage practice (source: Appendix table 1)

Fig 3.2 Parental marriage practice of TIPs (source: Appendix table 2)

Fig 3.3 Male TIPs marriage practice (source: Appendix table 3)

Fig 3.4 Female TIPs marriage practice (source: Appendix table 4)

Figure 3 TIPs marriage practice in circular layout by ethnic groups

First of all, TIPs marriage practice is visualized in Figure 3 in a circular layout of ethnic groups. It is worthy of noting that visualization method used in Figure 3 is widely used in genome sequencing studies of life sciences. It aims to demonstrate patterns and relationships between various types of ethnic marriage practice, including exogamy, intra-ethnic endogamy, and inter-ethnic endogamy. Ethnic groups in the circular layout include (1) the fifteen TIPs ethnic groups shown in Table 2 and (2) "non-TIPs" which refers to those whose ethnicity is not TIPs. Width of links represents volume of marriage practice. Links within same ethnicity refers to intra-ethnic endogamy, links between two different ethnic groups of TIPs represent inter-ethnic endogamy, and links between TIPs ethnic groups and "non-TIPs" represent exogamy.

Figure 3 demonstrates four visualized infographics of marriage practice, including TIPs marriage practice (Fig 3.1), parental marriage practice of TIPs (Fig 3.2), male TIPs marriage practice (Fig 3.3), and female TIPs marriage practice (Fig 3.4). Data that are applied to create the aforementioned four infographics are summarized in Appendix Table. As a whole, information revealed by Figure 3 suggests (1) that contemporary TIPs are associated with much more exogamy practices than their parents; this is particularly noteworthy for the ethnic Amis which serves as the largest ethnic group of TIPs; (2) that intra-ethnic endogamy are prevalent in the four largest ethnic groups; and (3) that female TIPs are associated with more exogamy practices than their male peers.

Table 1 demonstrates the extent to which individual characteristics such as gender, age, and education and spatial organization such as type of residential place are associated with practice of endogamy and exogamy. In terms of gender difference in choice between endogamy and exogamy, males are associated with higher share of endogamy practice than that of their female peers (68.7% for males and 58.2% for females). In other words, females are more likely than males to select exogamy practice. When it comes to the age effect on the choice between endogamy and exogamy, Table 1 shows that the share of endogamy practice with respect to different age groups exhibits a concave pattern. In other words, the share of endogamy is seen to decrease and thereafter increase with age, with the 35-44 age group having lowest share of endogamy practice. As for how education is associated with endogamy and exogamy, Table 1 clearly exhibits that the share of endogamy practice declines monotonically with educational level (73.6% for at-most primary education, 66.4% for junior/senior high, 59.7% for vocational high, 54.3% for some college, 45.6% for university, 41.1% for master degree, and 34.8% for doctoral degree). In other words, education has a negative (or positive) effect on the choice of endogamy practice (or exogamy).

It is also found that the practice of endogamy and exogamy has distinct pattern in different types of residential place. As shown in Table 1, the married who live in non-metropolitan areas are associated with a very high share of endogamy practice (71.8%) and thus a very low share of exogamy practice (28.2%). This situation is in strong contrast to their peers residing in metropolitan areas in which the corresponding share for endogamy practice and for exogamy practice is 46.0% and 54.0%, respectively. In short, marriage practices for the married living in non-metropolitan areas are much more associated with endogamy, whereas marriage practices for those living in metropolitan areas are mainly associated with exogamy.

We now move to highlight those whose marriage practice is endogamous. Table 2 demonstrates how individual characteristics such as gender, age, and education and attributes of residential place are associated with the choice between intra-ethnic and inter-ethnic endogamy for those whose marriage practice is endogamous. In terms of gender difference in choosing intra-ethnic and inter-ethnic endogamy, Table 2 shows that there is nearly no difference between males and females; in other words, there is not salient gender effect on the choice between intra-ethnic and inter-ethnic marriage for those with endogamous practice. For example, the share for intra-ethnic (or inter-ethnic) endogamy is of 80.9% (or 19.1%) for males and 81.0% (or 19.0%) for females. It is worthy of stressing that although there is no distinct gender effect on the choice between intra-ethnic and inter-ethnic endogamy, it is found that as aforementioned there is distinct gender effect on the choice between endogamy practice and exogamy practice.

Table 1 Endogamous and exogamous marriage practice for the married/spoused by sex, age, educational level, and type of residential place

Gender, age, education, residential place	The married & spoused (persons)*	Type of marriage practice		
		% Both	% Endogamy	% Exogamy
Overall	177,567 (100.0)	100.0	63.0	37.0
Gender				
Male	81,294 (45.8)	100.0	68.7	31.3
Female	96,273 (54.2)	100.0	58.2	41.8
Age (years)				
15-24	2,448 (1.4)	100.0	60.2	39.8
25-34	25,785 (14.5)	100.0	57.1	42.9
35-44	45,784 (25.8)	100.0	56.6	43.4
45-54	46,669 (26.3)	100.0	62.8	37.2
55-64	35,055 (19.7)	100.0	69.0	31.0
65+	21,826 (12.3)	100.0	74.7	25.3
Educational level				
Primary and less	41,685 (23.5)	100.0	73.6	26.4
Senior/Junior high	43,640 (24.6)	100.0	66.4	33.6
Vocational high	62,980 (35.5)	100.0	59.7	40.3
Some college	15,962 (9.0)	100.0	54.3	45.7
University	11,656 (6.6)	100.0	45.6	54.4
Master	1,598 (0.9)	100.0	41.1	58.9
Ph.D.	46 (0.0)	100.0	34.8	65.2
Residential place				
Non-Metro	117,296 (66.1)	100.0	71.8	28.2
Metro	60,271 (33.9)	100.0	46.0	54.0

Note data source: see note in Table 2

* those whose marital status is registered as either "married" or "spoused" in the source data

However, the age effect on the choice between intra-ethnic and inter-ethnic endogamy is very distinct. As shown in Table 2, the share of intra-ethnic endogamy increases monotonically with age; in other words, the share of inter-ethnic endogamy decreases with age. For example, the corresponding share of intra-ethnic endogamy is of 66.2% for those aged 15-24, 68.6% for those aged 25-34, 75.5% for those aged 35-44, 81.8% for those aged 45-54, 86.2% for those aged 55-64, 92.7% for those aged 65 and over, respectively. In effect, this finding suggests that age has a positive (or negative) effect on the choice of intra-ethnic endogamy (or inter-ethnic endogamy).

Education exhibits very distinct correlation with intra-ethnic and inter-ethnic marriage for those with endogamy practice. As suggested by Table 2, the share of intra-ethnic marriage monotonically declines with educational level; or in other words, the share of inter-ethnic marriage increases monotonically with education. For example, the share of intra-ethnic marriage is as high as 90.1% for those with primary education and less, 82.5% for senior and junior high, 76.6% for vocational high, 72.3% for education with some college, and 66.3% for university. It is worthy of highlighting that the educational effect on inter-ethnic marriage resembles the educational effect on exogamy practice indicated in

Table 1. When it comes to the correlation between attributes of residential place and type of endogamy practice, Table 2 demonstrates that for those with endogamy practice, the share for intra-ethnic marriage in metropolitan areas (81.8%) is not saliently different from the share in non-metropolitan areas (80.4%), suggesting that intra-ethnic marriage (or inter-ethnic marriage) is almost irrelevant to the metro/non-metro distinction. It is worthy of stressing that this finding is different from the finding demonstrated in Table 1 that metropolitan and non-metropolitan areas are associated with distinct difference in share of exogamy (or endogamy).

Table 2 Intra- and Inter-ethnic marriage practices for the married with endogamy by gender, age, educational level, and type of residential place

Gender, age, education, residential place	Individuals with endogamy practice (persons)*	Type of endogamy practice		
		% Both	% Intra-ethnic	% Inter-ethnic
Overall	111,910 (100.0)	100.0	80.9	19.1
Gender				
Male	55,838 (49.9)	100.0	80.9	19.1
Female	56,072 (50.1)	100.0	81.0	19.0
Age (years)				
15-24	1,474 (1.3)	100.0	66.2	33.8
25-34	14,729 (13.2)	100.0	68.6	31.4
35-44	25,924 (23.2)	100.0	75.5	24.5
45-54	29,288 (26.2)	100.0	81.8	18.2
55-64	24,191 (21.6)	100.0	86.2	13.8
65+	16,304 (14.6)	100.0	92.7	7.3
Education				
Primary and less	30,693 (27.4)	100.0	90.1	9.9
Senior/Junior high	28,973 (25.9)	100.0	82.5	17.5
Vocational high	37,595 (33.6)	100.0	76.6	23.4
Some college	8,661 (7.7)	100.0	72.3	27.7
University	5,315 (4.7)	100.0	66.3	33.7
Master	657 (0.6)	100.0	64.7	35.3
Ph.D.	16 (0.0)	100.0	43.8	56.3
Residential place				
Non-Metro	84,193 (75.2)	100.0	81.8	18.2
Metro	27,717 (24.8)	100.0	80.4	19.6

Note data source: see note in Table 2

* those whose marriage practice is endogamy as shown in Table 2

6 Conclusion and discussions

With gradual availability of massive micro data & substantial drop of digital hardware costs, computation for social complexity based on simplicity like the construction of micro genealogy becomes feasible. Computing issues remain challenging & total costs of computing are still time expensive. The emerging data science that integrates multidisciplinary skills & knowledge of “hacking skills”, “advanced math/stat”, and “domain knowledge” is crucial to overcome such constraint.

In order for the aforementioned variables being measurable, the research strives to construct a micro genealogy databank using high-performance computing methods to synthesize and extract information from various data sets. The constructed genealogy databank enables us to explore and examine not only the complex ethnic relationships at individual level, but also allows us to distinguish various types of ethnic marriage practice and construct the whole picture of ethnic identity formation. Thus construction of individual genealogy serves as the most important step to unveil the complex relationship between various marriage practices and ethnic identity formation.

TIPs are very ideal for studies on issues related to ethnic identity and ethnic marriage as well as social cohesion and wellbeing. Research findings based on TIPs turn out reasonable. The research at first explores patterns of ethnic marriage of TIPs, finding that they are associated with distinct pattern between parental and child generations, and between males and females. In addition to the effects of generation and gender, the research also finds that education and spatial organization are crucial in explaining ethnic marriage practice. The general findings include that parental marriage practice is mainly dominated by endogamy; that contemporary TIPs are associated with much more exogamy practices than their parents; that intra-ethnic endogamy are prevalent in the large ethnic groups; and that female TIPs are associated with more exogamy practices than their male peers.

The research finds that spatial organization matters. As a result of ethnic segregation and selectivity mechanism of migration process, the married who live in non-metropolitan areas are associated with a very high share of endogamy practice and thus a very low share of exogamy practice. This situation is in strong contrast to their peers residing in metropolitan areas. However, the research finds that intra-ethnic marriage (or inter-ethnic marriage) is almost irrelevant to the metro/non-metro distinction.

In the end, the research potential contribution lies in using methods embedded in the discipline of data science to enrich complex micro data sets which enable us to realize the hard-to-measure parental marriage practice and child ethnic identity formation. With gradual availability of massive micro data and substantial drop in costs of using advanced digital infrastructure, researches that utilize high-performance computing or even supercomputing to solve complex issues are no longer limited to the disciplines of natural and life sciences. In the past five year, we have seen the emergence of high-performance computing and supercomputing is reshaping research methods in social sciences. Computations for social complexity based on simplicity like the construction of micro genealogy in the research become feasible. Taking this research for example, computing issues remain challenging and total costs of computing are still time expensive. The emerging data science, that integrates multi-disciplinary skills and knowledge of manipulating digital infrastructure, expertise in advanced mathematics and statistics, and domain knowledge, is proven to be a crucial new way, or interdisciplinary study, that enables us to overcome some conventional research constraints.

References

Abdelal, R., Herrera, Y. M., Johnston, A. I., & McDermott, R. (2009). *Measuring Identity: A Guide for Social Scientists.* Cambridge University Press.

Alba, R., & Golden, R. (1986). Patterns of ethnic marriage in the United States. Social Forces, 65, 202–223.

Bellwood, P. (1991). The Austronesian dispersal and the origins of languages. *Scientific American*, 265(1), 88–93.

Blau, P., & Schwartz, J. (1984). *Crosscutting social circles.* New York: Academic press.

Blust, A. R. (1985). The Austronesian homeland: A linguistic perspective. *Asian Perspectives,* 26(1), 45–67.

Chan, J., To, H.-P. & Chan, E. (2006). Reconsidering social cohesion: Developing a definition and analytical framework for empirical research. *Social Indicators Research,* 75, 273–302.

Chandra, K. (2009). A constructivist dataset on ethnicity and institutions. In Rawi Abdelal, et al. (ed.s) Measuring Identity: A Guide for Social Scientists. Cambridge University Press.

Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council. 2013. Frontiers in Massive Data Analysis. Washington, D.C.: National Academy of Sciences.

Erikson, E. H. (1968). *Identity, youth, and crisis*. New York: Norton.

Gordon, M. (1964). *Assimilation in American life*. New York: Oxford University Press.

Gray, R. D., et al. (2009). Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science,* 323, 479–483.

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler. 2007. Data Quality and Record Linkage Techniques. Springer.

Hogg, M. A., Terry, D. J., & White, K. M. (1995). A tale of two theories: A critical comparison of identity theory with social identity source. *Social Psychology Quarterly*, 58(4), 255-269.

Hughes, D. (2003). “Correlates of African American and Latino parents’ messages to children about ethnicity and race: A comparative study of racial socialization”, *American Journal of Community Psychology,* 31, 15–33.

Hughes, D., Rodriguez, J., Smith, E. P., Johnson, D. J., Stevenson, H. C., Spicer P. (2006). Parents' ethnic-racial socialization practices: a review of research and directions for future study. *Developmental Psychology,* 42(5), 747-70.

Ji-Ping Lin. 2012. “Tradition and Progress: Taiwan’s Evolution Migration Reality 2012,” *Migration Information Source*, Washington D.C.: Migration Policy Institute.

Kalmijn, M. (1998). Intermarriage and homogamy: causes, patterns, trends. *Annual Review of Sociology,* 24(1), 395–421.

Koopmans, R. & Schaeffer, M. (2016). Statistical and perceived diversity and their impacts on neighborhood social cohesion in Germany, France and the Netherlands. *Social Indicators Research,* 125(3), 853-883.

Laurence, J. (2011). The effect of ethnic diversity and community disadvantage on social cohesion: A multi-level analysis of social capital and interethnic relations in UK communities. *European Sociological Review,* 27(1), 70-89.

Li, P, C.S Liu, I.H. Chang, and J.P. Lin. 2015. “Language shift of Taiwan’s indigenous peoples: a case study of Kanakanavu and Saaroa,” Journal of Multilingual and Multicultural Development. (forthcoming)

Lin, J. P. (2012). “Tradition and progress: Taiwan’s evolution migration reality 2012,” *Migration Information Source*. Washington D.C.: Migration Policy Institute.

Lin, J. P. (2013). “Are native "Flights" from immigration "Port of Entry" pushed by immigrants?: Evidence from Taiwan”,” in Fong, E, N. Chiang, and N. Delton (ed.s) Immigrant Adaptation in Multiethnic Cities - Canada, Taiwan, and the U.S , Rouledge.

Lin, J.-P., Tsai, B.-W., & Lee, M.-C. (2016, January 6). TIPD : Taiwan Indigenous Peoples open research Data 台灣原住民基礎開放研究資料庫. Retrieved from osf.io/e4rvz

Lin, Ji-Ping. 2013. “Are Native "Flights" from Immigration "Port of Entry" Pushed by Immigrants?: Evidence from Taiwan”,” in Fong, E, N. Chiang, and N. Delton (ed.s) *Immigrant Adaptation in Multiethnic Cities - Canada, Taiwan, and the U.S* , Rouledge.

Lin, Ji-Ping. 2013. “Casual Employment,” in Michalos, Alex (ed.) *Encyclopedia of Quality of Life Research*, Springer.

Lin, Ji-Ping. 2014. “Micro Discrete Events and Macro Continuous Social Outcomes: Migration Flows Analysis and Scientific Computing Challenges for Social Scientists,” in Proceedings of 2013 International Symposium on Grids & Clouds, PoS (Proceedings of Science). (Refereed).

McDoom, O. S. & Gisselquist, R. M. (2015). The measurement of ethnic and religious divisions: Spatial, temporal, and categorical dimensions with Evidence from Mindanao, the Philippines. *Social Indicators Research.* doi: 10.1007/s11205-015-1145-9

Monden, C., & Smits, J. (2005). Ethnic intermarriage in times of social change: The case of Latvia. *Demography*, 42(2), 23–345.

Nagel, J. (1994). Constructing ethnicity: Creating and recreating ethnic identity and culture. *Social Problems*, 41(1), 152-176.

O’Neil, Cathy and Rachel Schutt. 2014. Doing Data Science. CA: O’Reilly Media, Inc.

Phinney, J. S. & Ong, A. D. (2007). Conceptualization and measurement of ethnic identity: Current status and future directions. *Journal of Counseling Psychology*, 54(3), 271-281.

Phinney, J. S. (1990). Ethnic identity in adolescents and adults: A review of research. *Psychological Bulletin, 108*, 499–514.

Phinney, J. S., & Chavira, V. (1992). Ethnic identity and self-esteem: an exploratory longitudinal study. *Journal of Adolescence*, 15, 271-281.

Rochelle, T. L. (2015). Diversity and trust in Hong Kong: An examination of Tin Shui Wai, Hong Kong’s ‘City of Sadness’. *Social Indicators Research*, 120, 437–454.

Schmidt, Eric and Jared Cohen. 2013. The New Digital Age: Reshaping the Future of People, Nations and Business. 1st Edition. Knopf Inc.

Schwartz, S. J. (2001). The evolution of Eriksonian and neo-Eriksonian identity theory and research: A review and integration. *Identity*, 1, 7–58.

Schwartz, S. J., Zamboanga, B. L., Wang, W., & Olthuis, J. V. (2009). Measuring identity from an Eriksonian perspective: Two sides of the same coin? *Journal of Personality Assessment*, 91(2), 143–154.

Shutler, R. Jr. & Marck, J. C. (1975). On the dispersal of the Austronesian horticulturalists. *Archaeology and Physical Anthropology in Oceania*, 10(2), 81–113.

Smith, E. J. (1991). Ethnic identity development: Toward the development of a theory within the context of majority/minority status. *Journal of Counseling and Development*, 70, 181-188.

Smith, E. P., Walker, K., Fields, L., Brookins, C. C., & Seat, R. C. (1999). Ethnic identity and its relationship to self-esteem, perceived efficacy and prosocial attitudes in early adolescence. *Journal of Adolescence*, 22, 867-880.

Smits, J. (2010). Ethnic intermarriage and social cohesion. What can we learn from Yugoslavia? *Social Indicators Research,* 96, 417–432.

Spencer, M. B. & Markstrom-Adams, C. (1990). “Identity processes among racial and ethnic minority children in America”. *Child Development,* 61, 290–310.

Stryker, Sheldon. & Burke, Peter J. (2000). The past, present, and future of an identity theory. *Social Psychology Quarterly,* 63(4), 284-297.

Tajfel, H. (1981). *Human groups and social categories.* Cambridge, England: Cambridge University Press.

Turner, J. C., Hogg, M. A., Oakes, P. J., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford: Blackwell.

Verkuyten, M. & Thijs, J. (2004). Global and ethnic self-esteem in school context: minority and majority groups in the Netherlands. *Social Indicators Research*, 67, 253–281.

Wood, David (Ed.). 2011. Linking Government Data. Springer.