Lecture LIV
TIGER DATA SET

There is usually a gap between the real world data formats and the data structures which are theoretically the most appropriate for algorithms. Witness the plethora of graphics file formats, and the many trade books that treats them. In this lecture, we will describe an important format called TIGER (Topologically Integrated Geographic Encoding and Referencing System) developed by the U.S. Bureau of the Census (http://www.census.gov). For more information, we recommend going to the TIGER homepage: http://www.census.gov/geo/www/tiger.

§1. Overview

The TIGER data set contains the geographic encoding of the whole USA, that is especially appropriate for representing census information, its primary application. For our purposes, we are mainly interested in its map data.

TIGER files. The data set is county-based, so that each county in the county has its own data set. The whole USA has over 3200 counties. Each county is given a 6 digit identifier called an FIPS. For instance, the FIPS for New York County (i.e., Manhattan) is 36061 while Kings County (i.e., Brooklyn) is 36047. The first two digits of the FIPS identifies the state (New York State is 36).

Each county has up to 17 different record types, and all the records of a given type for the county is stored in a single file. So, there are up to 17 files for each county. The files for Manhattan are named TGR36061.RT1, TGR36061.RT2, TGR36061.RTA, etc. The suffix RT1, RT2, etc, tells us that the file contains "Record Type 1", "Record Type 2", etc. The file suffixes are RTn where n is one of the characters

$$1, 2, \ldots, 9, A, C, H, I, P, R, S, Z.$$

Each file is an ASCII file. Each record is stored in one line of the file, and each line has a fixed number of ASCII characters. Each record also has a fixed number of fields, and fields occupy predetermined positions in its line. Below, we give more details about the contents of these files.

Coordinate System and Accuracy. The coordinate system is based on Longitude and Latitudes. Each Longitude is given as a signed 9 digit sequence, with an implied 6 decimal places. Thus -123456789 really represents -123.456789. Each latitude is given as a signed 8 digit sequence, also with an implied 6 decimal places. Thus +12345678 really represents +123.45678. For instance, the bounding box for New York State is (minLon, maxLon, minLat, maxLat) = (-79.762418, -71.778137, 40.477408, 45.010840).

You will need to make a conversion from the Tiger coordinates to conventional Lat/Lon coordinates which are specified in degree and minutes. Moreover, latitudes are relative to the Equator (North or South) and longitudes are relative to Greenwich Meridian (East or West). For instance, the Tiger coordinates for Courant Institute (NYU) is roughly (40.729, -73.995). This converts to (40 deg 43.74' North, 73 deg 59.70' West). Thus, negative coordinates are South of Equator or West of Greenwich Meridian.

The accuracy of the map is that of a 1:100,000 scale map, which is correct up to ±167 feet. However, the relative positions of any specified points (with respect to the plane subdivision of the Tiger data) is correct.

TIGER Geometry. There are three kinds geometric objects: points, polygonal lines and polygons. The polygonal lines in TIGER files have the registered name of TIGER/Line (R). They are also known as "complete chains", but in this lecture, we simply call them **Tiger Lines**. Each Tiger Line has a unique ID (or TLID). Each polygon, which we will call **Tiger Polygons**, also has an ID (or POLYID), but this ID is unique only within each county. But is easy to make a POLYID unique across the whole country by concantenating it with the FIPS, for instance. Each Tiger Line is basically the maximal polygonal chain that share the same "line features". We will discuss line features below, but the type of a line feature might be "road" or "county boundary". The basic Tiger Line information is stored in RT1 and RT2.

Points (or vertices) in a Tiger line are of two kinds: of Tiger Lines, and non-endpoints. The latter are called **detail points**. Every Tiger Line has two endpoints, and zero or more detail points. In the Tiger File RT1, the endpoints are stored with each Tiger Line. The detail points are stored in RT2. Since Tiger files have fixed size records, while the number of detail points for a Tiger Line has no a priori bound, the solution is to group the detail points for a Tiger Line into groups of ten detail points per record in RT2 files. At most one group has less than 10 detail points. If a Tiger Line has no detail points, then it has no corresponding entry in the RT2 file.

It is important to note that points in the Tiger Line do not have independent existence (i.e., no unique ID's). This could potentially lead to inconsistencies, both within a county and across two adjacent counties. Inconsistencies cannot happen with detail points, as they are not shared with other Tiger Lines. But endpoints will generally be shared by two or more Tiger Lines. If two Tiger Lines L, L' share an endpoint P, then the coordinates of P in L and in L' must agree. This consistency is not encoded into the Tiger data organization, but is an implicit guarantee.

The set of Tiger Polygons forms a subdivision of its county (and by extension to the whole USA). Each Tiger Polygon is associated with an interior point. The boundary of each Tiger Polygon is bounded by a whole number of Tiger Lines; this implies that each Tiger Line bounds one or two Tiger Polygons. A Tiger Lines bounds only one Tiger Polygon if the other side of the line is outside the county or outside the map coverage.

§2. Tiger Record Types

Fix any county. Each Record Type (RT) is stored in a single file. To understand the basic information in these files, we use some basic concepts from Relational Database theory. Each file is thus a Relation. A relation has a fixed set of attributes. We suggest two steps to figuring out the Tiger files:

(1) To understand an individual file, you need to figure out that one of its attributes is the "KEY". E.g., in RT1, the TLID field is the key. This KEY is unique within the current county in the sense of relational data bases. With the exception of TLID, most keys are not unique across counties or states (so they need to be combined with some other attributes). So, the first thing is understand for each file to identify its KEY.

(2) Second, to connect the information across files, you need to find shared attributes. For instance, the POLYID attribute is found in RT9 and RTA. By "joining" these two files in the sense of relational database, you can cross reference properties.

Note that blank entries are very common, and it means that the corresponding attribute is not applicable for that row. Alternatively, you might say that the database scheme for these relations

The table below lists some important fields in the various Tiger files:

Record Type	Field Name	Field Position	Description
RT 1 (Basic info for Lines)	TLID	6-15	TIGER/Line ID
	SIDE1	16	Single side? (blank = both sides, 1 = single side)
	FRLONG	191-200	Start (from) longtitude
	FRLAT	201-209	Start (from) latitude
	TOLONG	210-219	End (to) longtitude
	TOLAT	220-228	End (to) latitude
RT 2 (Detail points for Lines)	TLID	16-25	TIGER/Line ID (not all lines have an entry here
	RTSQ	16-18	Record Sequence Number (1, 2, etc)
	LONG1	19-28	Start (from) longtitude of point 1
	LAT1	29-37	Start (from) latitude of point 1
	LONG2	28-37	Start (from) longtitude of point 2
	LAT2	etc	etc
RT A (Basic info for Polygons)	POLYID	6-15	Polygon ID
RT I (Line-Polygon info)	TLID	6-15	TIGER/Line ID
	POLYIDL	27-36	Polygon ID on the left side
	POLYIDR	42-51	Polygon ID on the right side
RT P (More info on Polygons)	POLYID	16-25	Polygon ID
	POLYLONG	26-35	Internal Point Longtitude
	POLYLAT	36-44	Internal Point Latitude

N.B. The html version of this table may not be completely correct; refer to the postscript version.

Example: Metropolitan Areas. Suppose you are interested in Metropolitan Areas (MA's). This information is found in Record Type S (RTS). The KEY of RTS is POLYID (polygons). Two attributes found in RTS are MSA/CMSA and PMSA. What are these? Well, there are two basic kinds of MA's: Metropolitan Statistical Areas or MSA (under 1 million population), Consolidated MSAs or CMSA (over 1 million population). CMSA's are in turn subdivided into Primary MSAs or PMSAs. E.g., New York City is a CMSA, but it has many PMSA's. Thus, to identify an MA, you look for entries in the MSA/CMSA attribute. If a row (i.e., a polygon) is part of an MSA, then its PMSA attribute is blank; if it is a CMSA, it will have a PMSA attribute.

Non-Geometric and Non-topological Data. Such data include census data (of course), postal address ranges, zip code, land type and metropolitan areas. Also of practical interest is landmarks (school, park, airport, etc).

Exercises

Exercise 2.1: Suppose you are given a polygon ID.
(a) How do you check if that polygon represents water or land?
(b) How do you check if that polygon represents some interesting landmark, and determine the name of that landmark? ◇

Exercise 2.2: Determine the KEY attribute in each Record Type. ◇

Exercise 2.3: (a) Write a simple program to display the formatted information in any Tiger file (in the style of a spread sheet). The formatting of the file should be stored in a separate "tiger.style" file which can be editted. The most basic information for the style file are the column names and start/end positions.
(b) Generalize the above style file to allow translation. For instance, if a value of "1" or "0" means "water" or "land", we want to be able to tell the display to show "W" or "L" instead of "1" or "0". Various other codes can be similarly translated. $\diamondsuit$

Exercise 2.4: Outline what processing steps are needed to create a map of the US which picks out all its Metropolitan Areas (say, shaded in pink). Which Record Types must be used here? $\diamondsuit$

Exercise 2.5: Describe a method to locate all the streets in a county whose name begins with "Wash" (e.g., Washington, Washburn, etc). $\diamondsuit$

Exercise 2.6: We would like to be able to draw Tiger maps with information about location of towns and metropolitan areas. We want to represent these towns by "red dots". Experimentally determine answers to the following:
(i) At what scales should these red dots appear?
(ii) How do you extract the location and names of these red dots from the Tiger data set? HINT: in Record Type C, there is a list of "geographic entities", which could be townships. In Record Type A, each polygon is associated with up to three geographic entities.
(iii) Discuss any related issues. $\diamondsuit$

End Exercises

§3. Issues

Of course, the conversion issue is to extract from the Tiger dataset the usually connectivity expected in structures such as half-edge data structures. What makes this interesting is that we may want to process this incrementally (such as across the netwok).

When we display maps based a global referencing system (usually this means the longitude/latitude system), it is seldom acceptable to treat this as if the coordinates come from a rectangular coordinate system. Otherwise, the map would look distorted. We need to choose some map projection scheme. That is, if G is the surface of the spherical globe, we want to choose a projection surface P and a partial map $\pi : G \to P$. The simplest projection is the **sterographic projection**. In this case, P is a plane tangent to G at any chosen point $p_0 \in G$. For any point $p \in G$, we define $\pi(p) \in S$ to be the intersection of S with the ray $\rho(p)$ that emanates from the center of the globe and passing through p. Note that $\pi(p)$ is undefined for points in a hemisphere and $\pi(p_0) = p_0$.

Another simple projection is to take P to be cylinder that touches G at the equator. Again, the projection point $\pi(p)$ is defined to the intersection of $\rho(p)$ with P. In this case, $\pi(p)$ is undefined for only two points, the North and South poles. This **cylindrical projection** is is easy to calculate if $p \in G$ is given in terms of longitude-latitude $(\theta, \phi) \in [-\pi, \pi) \times (-\pi/2, \pi/2)$. and the cylinder P has the rectangular coordinates $[-\pi, \pi) \times \mathbb{R}$. Assume that points (θ, ϕ) on the equator projects to $\pi(\theta, \phi) = (\theta, 0)$. We leave this as an exercise.

When we merge counties, there may be inconsistencies or unexpected topology. For instance, we discovered in merging Manhattan with neighboring New Jersey counties that the Liberty Island belongs to

Manhattan County but is surrounded by water belonging to New Jersey. How do we check for topological consistency and regularity?

Exercises

Exercise 3.1: Write a Java program to pull out all Tiger Lines and Polygons that lie within some longitude/latitute range. ♢

Exercise 3.2: Derive formulae for the projections discussed above:
(a) stereographic projection with respect to a point p_0,
(b) cylindrical projection. ♢

Exercise 3.3: (a) Write a program which, given a computer file directory (storing all the Tiger files for a single county), will extract the minimum bounding box (in terms o longitude/latitude) for the county.
(b) Construct a data base for the entire USA storing a minimum bounding box for each county.
(c) Using this, write a program to construct a list of all the counties that (potentially) intersects a given longitude/latitute range. ♢

Exercise 3.4: Write a program for checking the topological consistency of the data in (a) a single Tiger county, and (b) in several Tiger counties. Basically we need to check that the endpoints of Tiger Lines are consistent. ♢

Exercise 3.5: Write a program which, given a directory storing the Tiger files for a single county, and a postal address, will compute and determine the longitude/latitude of a point near this address. Extra credit: also extract the zip code. ♢

§4. HUMAN INTEREST DATA

In the above description of the Tiger data, we focused on the geometric information (lines and polygons, lat/long, adjacency relations, etc). But of human interest are the names and types of various features, location of landmarks, etc. Maps would be useless without this information.

Coloring of Map Areas. The first basic data in any map is the distinction between water (blue) and land. For land, we want to color code park areas (green) and public facilities or institutions such as airports and hospitals (pink), and the rest (yellow). Record Type S has a water flag for polygons.

Landmarks. Record Type 7 contain landmark features. Each landmark is given a ID (positions 11-20), name (positions 25-54) and longitude/latitude (positions 55-64/65-73). There is also a **census feature class code** (CFCC) for each landmark. This is a 3 byte code in positions 22-24. This information is critical in our map color coding.

Some landmarks are area landmarks (like parks and hospitals). Others are point landmarks (for instance a building). For area landmarks, their associated polygon (POLYID) can be found in Record Type 8.

Key Geographical Locations (KGL). This is found in Record Type 9. Example of a KGL are public squares and plazas. Each KGL has two associated fields POLYID and FEAT, which can be used to locate it on the map. To get to the POLYID location, use RT I, and to get to the FEAT location, use RT 5.

Line Features. A "line feature" is an informal concept that is typically a sequence of one or more contiguous Tiger Lines that share common attributes such as **feature identifiers**. For instance, a street named "Broadway" is a line feature that may be decomposed into many Tiger Lines. The Tiger data set does not guarantee that line features can be reconstructed easily from its data.

Each Tiger Line has a feature identifier in file RT1, stored in the field positions 18 to 55. Example of a feature identifiers: **N Adams Av**, **US Highway 1**, **Jefferson St**, **Providence St NE**. Feature identifiers are subdivided into 4 subfields:

- Feature Direction, Prefix (FEDIRP), in positions 18-19. E.g., **N** in "N Adams Av". Possible values: N, S, E, W, NE, NW, SE, SW, EX (for extended or extension).
- Feature Name (FENAME), in positions 20-49. E.g., **Adams** in "N Adams Av", **1** in "US Highway 1".
- Feature Type (FETYPE), in positions 50-53. E.g., **Av** in "N Adams Av", **US Highway** in "US Highway 1". Possible values are Av (or Ave), St, Rd, Dr, Ln, Aly Ct, Blvd, Way, Pky, Fwy, Hwy, Plz, Cir, Sq, Brg, Pl, Tpke, Ramp, Ter, Walk, Cres, Mal, Tunl.
- Feature Direction, Suffice (FEDIRS), in positions 54-55. E.g., **NE** in "Providence St NE".

However, each Tiger Line also has a **census feature class code** (CFCC) in positions 56-58. This code is highly correlated with the Feature Type (FETYPE), but are independently assigned.

Zip Code and Address Range. The zip code of the polygons to the left and right a each Tiger Line is stored in positions 107-116 of the RT1 records. The address range on the left and right side of each street is also recorded in positions 59-106.

CFCC Code. Such codes are found in record types RT1 (for Tiger Lines), RT7 (for point or area landmarks), RT9 (for key geographic location). Here is an illustrative list of the codes.

Class A code is for roads:

- Primary Highway with Limited Access such as interstate highways. Unseparated (A11), unseparated in tunnel (A12), unseparated in underpass (A13), etc.
- Primary Highway without Limited Access such as US highways, state and county highways. Unseparated (A21), etc.
- Secondary and Connecting Roads, such as those connecting small towns. A31, etc.

Class B code are for rail roads. Class C code are for miscellaneous ground transportation. E.g., C10 is pipeline, C20 is power transmission line. Class D code for landmarks:

- Military installations (D10)

- Multihousehold Quarters: hotel, motel, YMCA, YWCA (D27), campground (D28), shelter (D29)
- Custodial facility: hospital (D31), jails (D36)
- Educational or religious institution: general (D40), educational (D43), religious (D44)
- Transportation: general (D50), airport or airfield (D51), train station (D52), bus terminal (D53), marine terminal (D54).
- Employment Center: general (D60), shopping or retail center (D61), industrial building or park (D62), office building or park (D63), amusement center (D64), government center (D66).
- Tower: general (D70), lookout tower (D71).
- Open Space: general (D80), cemetary (D81), national park (D82), national forest or federal land (D83), state or local park or land (D85).
- Special purpose: general (D90), fire department (D93), library (D95), city/town hall (D96)

Class E is for physical features:

- Fence (E10).
- Topographic feature: general (E20), ridge line (E21), mountain peak (E22), island with name (E23), levee or embankment (E24).

Class F is for nonvisible features such as property areas, legal and administrative entities. These are only identified if they do not follow visible features such as roads or streams.

Class G is for US Census Bureau internal usage. Class H is hydrography.

- Basic category is for all shorelines regardless of the classification of the water: general (H00), perennial water (H01), intermitten water (H02).
- Natural Flowing Water: perennial stream or river (H11), intermittent (H12).
- Man-made channel to transport water (H21, H22).
- Natural Inland body of water: general (H30), perennial lake or pond (H31), intermittent (H32).
- Man-made body of water: perennial reservoir (H41), intermittent (H42).
- Seaward body of water: bay (H51), sea or ocean (H53)

Feature Class P, Provisional Features, is a new class that that may only appear on street features. These features are treated exactly line Class A.

EXERCISES

Exercise 4.1: Assume there is a global variable identifying the "current county". Write the following functions which looks in the current county:
(1) LandmarkID("NAME") which returns the ID of the landmark whose name matches "NAME". E.g., for Manhattan, LandmarkID("city hall") is ???
(2) LandmarkBB("LAND") which returns the bounding box containing the landmark whose ID is LAND ♢

Exercise 4.2: (1) Outline an algorithm to find a reasonably good path between any two specified street addresses within one county. Assume only the TIGER files, without any additional processing.
(2) What additional processing would be helpful in (1)?
(3) Sketch additional processing and data structure support needed if the addresses belong to different counties. ♢

Exercise 4.3: Map Visualization Project. Construct a complete binary tree whose leaves are the following 16 counties in the metropolitan NYC area:

No.	FIPS	Name
1.	36103	Suffolk
2.	36059	Nassau
3.	36081	Queens
4.	36047	King (Brooklyn)
5.	36061	New York (Manhattan)
6.	36085	Richmond (Staten Is)
7.	36005	Bronx
8.	36087	Rockland
9.	36071	Orange
10.	36119	Westchester
11.	34003	Bergen
12.	34017	Hudson
13.	34013	Essex
14.	34039	Union
15.	34023	Middlesex
16.	34025	Monmouth

The tree is supposed to be a "merge tree". Using specified in-order listing of the counties will to ensure that each internal node represents a connected geographical region. We want to introduce two operators: merge and simplify to construct a hierarchical map of the 16 counties. Merging amounts to removing common boundary data. Simplification amounts to discarding details. You also need to construct a viewer to visualize this data that you have constructed. We suggest using OpenGL or Java for this project. ♢

END EXERCISES