Arriving On Time? Finding Reliable Shortest Paths in a Stochastic Network

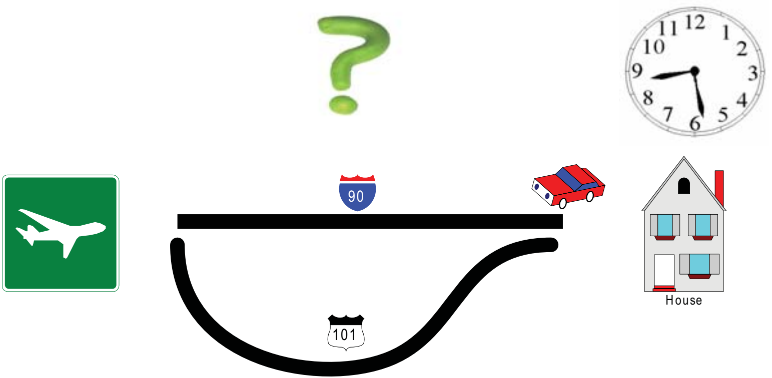
Marco Nie and Xing Wu [1]
John Dillenburg and Peter Nelson [2]

[1]Northwestern University
[2]University of Illinois, Chicago

CATMUG Monthly Seminar, April 7th, 2010

Outline

Motivation



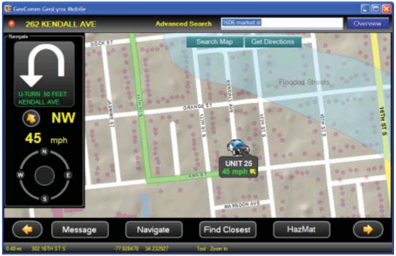
Question

When should you leave home, and which route should you take, if you need to drive to an important appointment, such as catching a flight or a job interview?

Motivation

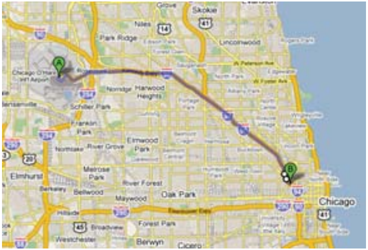
Google Map Driving Direction?

Built-in or adds-on navigation system?

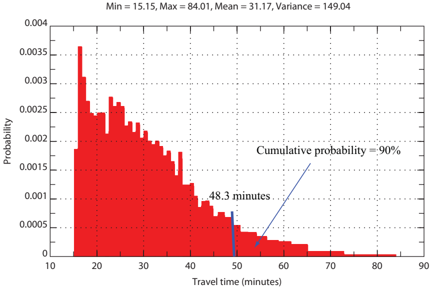


- Perhaps using driving direction provided by Google, Yahoo or your in-vehicle navigation system?
- Do you really trust their estimation of travel time when you don't want to miss that appointment?

Motivation



(a) Interstate 94/90 from Chicago (Ohio St.) to Ohare International Airport (source: Google Map)



(b) Travel Time Distribution for that corridor during morning rush hour (6-10 AM)

- Travel times vary from as low as about 15 minutes to as long as 80 minutes in the morning peak period (6 - 10 AM).
- If a traveler wishes to capture the flight on time with a 90% chance, 48 minutes have to be budgeted for travel, over 50% more than the mean travel time (31 minutes).

Considering travel reliability is important...

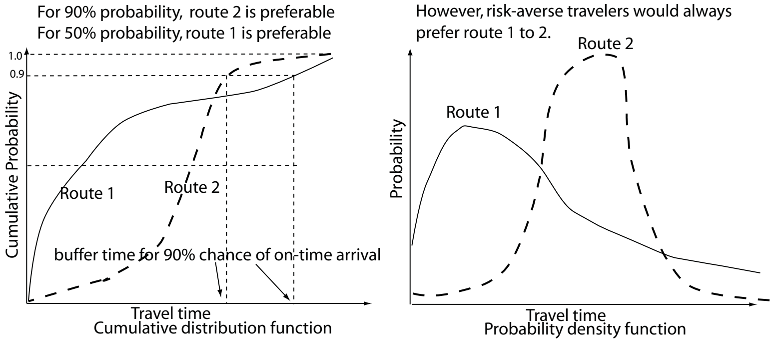
- Travelers need to incorporate reliability into route choice so that they can better use their time;
- Shippers and freight carriers need predictable travel times to fulfill on-time deliveries in order to remain competitive;
- The ability to arrive on-time with high reliability is imperative to emergency responders;
- Planning agency need to anticipate travelers' response to reliability in their planning process;
- ...

Reliable a priori shortest path problem (RASP) often arises from these applications

Problem statement

Problem

Assume: analytical or empirical probabilistic distributions of travel times on all roads are known;
Find: optimal *a priori* paths that require smallest time budget to ensure arriving on-time or earlier for a desired likelihood.



Stochastic routing problem

Minimize expectation

The basic problem is trivial, but complexity is introduced when the following issues are considered.

- Time-dependent networks: Hall (1986a), Fu (2001), Miller-hooks (2001), Fu & Rilett 1998, Miller-hooks & Mahmassani 2000.
- Correlated distributions: Waller & Ziliaskopoulos (2002), Fan et al. (2005b)
- Recourse: Croucher (1978), Andreatta & Romeo (1988),Polychronopoulos & Tsitsiklis (1996), Waller & Ziliaskopoulos (2002), Provan (2003), Gao & Chabini (2006).

Literature (cont.)

Maximize reliability

- Maximize the probability of realizing a travel time equal to or less than a given threshold: Frank (1969), Mirchandani (1976), Fan et al. (2005a), Nie and Wu (2009a,b,c).
- Maximize the probability of being the shortest: Sigal et al. (1980)
- Least possible travel time: Miller-hooks & Mahmassani (1998)
- Maximize expected utility: Loui (1983),Eiger et al. (1985), Murthy & Sarkar (1998)
- Minimize the maximum travel time: Yu & Yang (1998), Montemani & Gambardella (2004)

Literature (cont.)

Maximize reliability

- Maximize the probability of realizing a travel time equal to or less than a given threshold: Frank (1969), Mirchandani (1976), Fan et al. (2005a), Nie and Wu (2009a,b,c).
- Maximize the probability of being the shortest: Sigal et al. (1980)
- Least possible travel time: Miller-hooks & Mahmassani (1998)
- Maximize expected utility: Loui (1983),Eiger et al. (1985), Murthy & Sarkar (1998)
- Minimize the maximum travel time: Yu & Yang (1998), Montemani & Gambardella (2004)

Setting

Notation

- Consider a directed network $G(\mathcal{N}, \mathcal{A}, \mathcal{P})$ consisting a set of nodes $\mathcal{N}$ ($|\mathcal{N}| = n$), a set of links $\mathcal{A}$ ($|\mathcal{A}| = m$), a probability distribution $\mathcal{P}$ describing the statistics of the link traversal times (or costs).
- The traversal times of link ij (denoted as c_{ij}) is an independent random variable, following a given distribution $p_{ij}(\cdot)$.
- Travel time on path k^{rs} (which connects node r and the destination s) is denoted as π_k^{rs} and all paths that connect r and s forms a set of K^{rs}.
- The destination of routing is denoted as s.

Define optimality

Definition (b-reliable path)

A path k^{rs} is said b-reliable if and only if $u_k^{rs}(b) \geq u_l^{rs}(b), \forall l^{rs} \in K^{rs}$, where $u_k^{rs} = P(\pi_k^{rs} \leq b)$ denotes the cumulative distribution function (CDF) of π_k^{rs}.

Problem statement

A b-reliable path is the path that is most reliable with respect to b. Our goal is to find such reliable paths for every b.
However, dynamic programming is not directly applicable because

Theorem

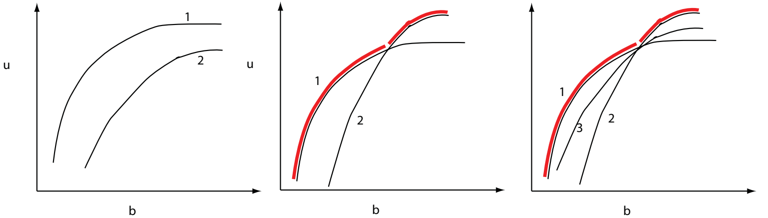
Subpaths of a b-reliable path may not be b-reliable.

First-order stochastic dominance (FSD)

Definition (FSD-admissible path)

A path $k^{rs} \in K^{rs}$ is FSD-admissible if and only if $\exists$ no path $l^{rs} \in K^{rs}$ such that 1) $u_l^{rs}(b) \geq u_k^{rs}(b), \forall b$, and 2)$\exists$ at least one b such that $u_l^{rs}(b) > u_k^{rs}(b)$.

FSD-admissible paths can be understood as non-dominant paths.



Path 1 is FSD-admissible
Path 2 is not. It is domiated by 1
Path 1 forms the pareto frontier

Both Path 1 and 2 are admissible
They together form the pareto frontier

All three paths are FSD-admissible
Path 3 does not contribute to the frontier, but it is not dominated by either 1 or 2.

Two results

Theorem

Subpaths of any FSD-admissible path must be FSD-admissible.

Two results

Theorem

Subpaths of any FSD-admissible path must be FSD-admissible.

- We can still search FSD-admissible paths using dynamic programming
- We have to deal with a set of such paths, which could grow exponentially with problem size.

Two results

Theorem

Subpaths of any FSD-admissible path must be FSD-admissible.

- We can still search FSD-admissible paths using dynamic programming
- We have to deal with a set of such paths, which could grow exponentially with problem size.

Theorem

A FSD-admissible path is acyclic.

Two results

Theorem

Subpaths of any FSD-admissible path must be FSD-admissible.

- We can still search FSD-admissible paths using dynamic programming
- We have to deal with a set of such paths, which could grow exponentially with problem size.

Theorem

A FSD-admissible path is acyclic.

- We can ignore paths with cycles
- This fact may be used to improve computational efficiency.

Solution procedure

Label-correcting

- Step 0: Initialization. Add a path starting and ending at the destination s into candidate list Q.
- Step 1: If Q is not empty, take a path k^{js} from Q, go to step 2; otherwise terminate.
- Step 2: For each path $k^{is} = ij \diamond k^{js}$, if it is FSD admissible, add it into Q, and remove all existing paths dominated by this k^{is}. Go back to Step 1.

Theorem (Finite convergence)

The above procedure terminates after a finite number of steps and yields a set of FSD-admissble paths for each node i.

Complexity

Bad news

The algorithm is non-deterministic polynomial, because the number of FSD-admissible paths may grow exponentially with the network size. The algorithm runs in order of $O(mn^{2n-1}L + mn^nL^2)$.

Complexity

Bad news

The algorithm is non-deterministic polynomial, because the number of FSD-admissible paths may grow exponentially with the network size. The algorithm runs in order of $O(mn^{2n-1}L + mn^nL^2)$.

Good news

- $|K^{is}|$ is much smaller than n^{n-1} for sparse networks commonly seen in transportation applications.
- The expected number of FSD-admissible paths is bounded roughly by $\log(|K^{is}|)$ if the number of discrete time points L is 2.

Complexity (cont.)

What if $L > 2$?

Get a theoretical bound is more difficult. However, through experiments we conjecture

- The number of FSD-admissible paths increases exponentially with L in general, and
- Due to the monotonicity of CDF, it seems to be bounded by $L \log(|K^{is}|)$.

If the second conjecture is correct, we can push the complexity to $O(mn^2L^3(\log(n))^2)$. This is a pseudo-polynomial bound!

Implementation issues

Extreme-dominance approximation

- Ignore FSD-admissible paths that do not contribute to the frontier
- The complexity of the solution procedure is now in the order of $O(mnL + mL^3)$ ($\simeq O(mL^3)$).
- This approximation does not always yield correct Pareto-frontiers.

Cycle avoidance

- A path with cycles cannot be FSD-admissible.
- It is thus useful to prevent paths with cycles from entering the current path set. The cost of such operations is well paid off.

Implementation issues (cont.)

Convolution integral

- The single most time-consuming component in the algorithm.
- Adaptive discretization schemes. The number of support points is bounded from the above, and is allowed to vary according to the shape of probability density function. The adaptive scheme achieves a satisfactory balance of efficiency and accuracy (Nie et al. 2010).
- Fast Fourier Transformation (FFT) can be used to further expedite the operation. It will reduce the quadratic complexity (L^2) to a logarithm one ($L \log L$). However, FFT is is effective only when L is relatively large ($> 10,000$).

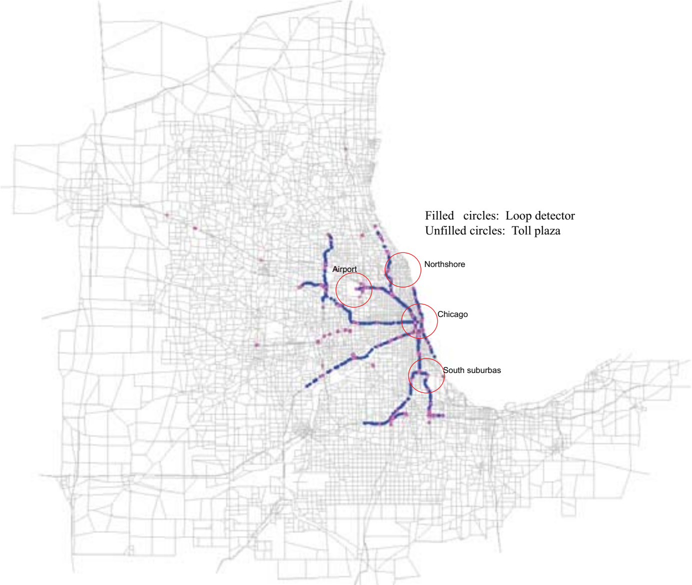
Chicago metropolitan region

- The third largest metropolitan area in the US and one of the most congested cities.
- The travel time in the Chicago area is more unreliable than any other major metropolitan areas in the US (planning index = 2.07, Mobility Report 2007).
- Chicago has archived a rich set of traffic data in both public and private sectors

Data

GCM (Gary-Chicago-Milwaukee corridor) traveler information system (www.gcmtravel.com) provide traffic data collected from loop detectors and electronic toll transponders (known as I-PASS).

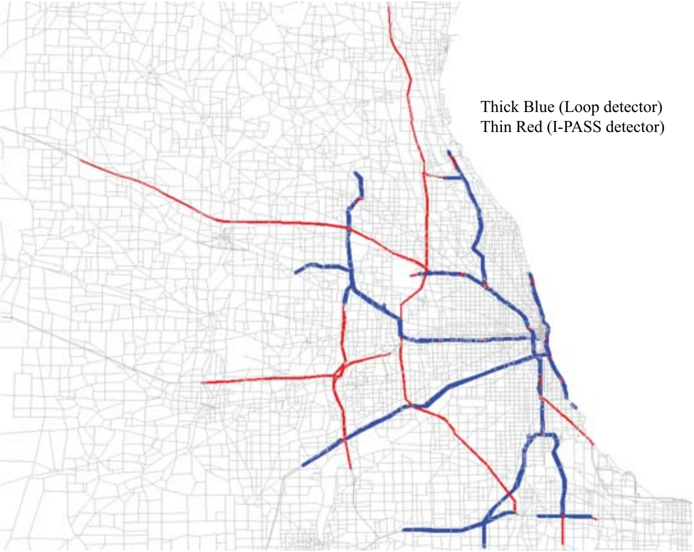
An overview of Chicago network



Data on freeway and toll roads

- Loop detectors record speed, occupancy and flow rate approximately every 5 minutes
- Travel times on toll roads between two I-PASS toll booths are obtained from in-vehicle transponders and aggregated every 5 minutes.
- About 825 loop detectors and 174 I-PASS detectors from GCM database are used.
- The loop detector data collected from 2004 10/10 to 2008 10/11, and the I-PASS detector data from 2004 10/9 to 2008 7/3.
- In total, 765 links are “covered" by either I-PASS detector, loop detector, or both.

Data coverage



Thick Blue (Loop detector)
Thin Red (I-PASS detector)

Construct distributions for covered links

Procedure

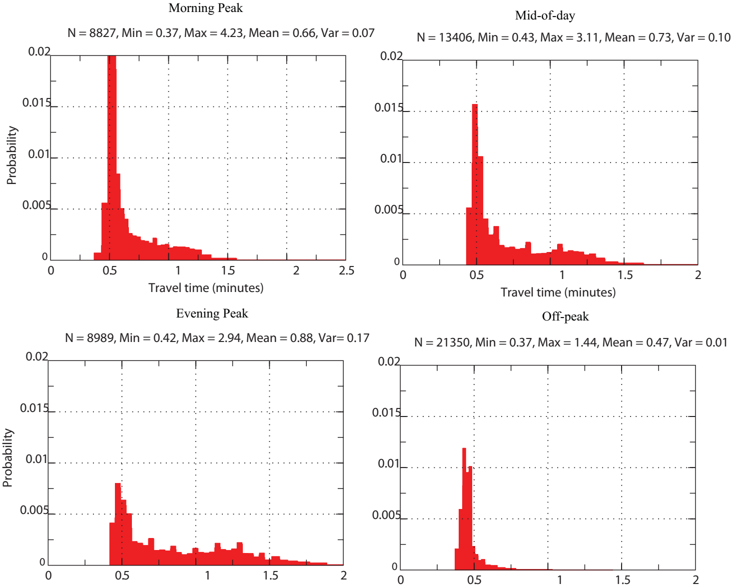
Step 1 Find $L_a = \min\{\tau_a(t), \forall t \in \Lambda\}, U_a = \min\{10l_a/v_a^0, \max\{\tau_a(t), \forall t\}\}$, where Λ is a set of valid time intervals in the observation period, and v_a^0 is free flow speed (or speed limit) on link a.

Step 2 Divide $[L_a, U_a]$ into M intervals, and let $\delta_a = (U_a - L_a)/M$. Find the set $D_m = \{\tau_a(t) | \forall t \in \Lambda, (m-1)\delta a \leq \tau_a(t) < m\delta\}, \forall m = 1, ..., M$

Step 3 Obtain the probability mass for each interval m using $P_m = \frac{|D_m|}{|\Lambda|}$.

The data are disaggregated into 150 different groups based on three factors: time of day (4 + 1), day of week (5 + 1) and season (4 + 1). Each covered link has 150 different distributions.

Sample distribution for different time of day



Data on arterial streets

Two step estimation process

The travel time distributions on arterial streets have to be estimated indirectly because no observations are available.

- Select an appropriate functional form: travel time on freeway and arterial is known to closely follow a Gamma distribution
- Estimate mean and variance

The probability density function of a Gamma distribution is

$$f(x) = \frac{1}{\theta^{\kappa}\Gamma(\kappa)}(x-\mu)^{\kappa-1}e^{-(x-\mu)/\theta}; x \geq \mu, \theta, \kappa \geq 0 \qquad (1)$$

where θ is the *scale parameter*; κ is the *shape parameter*; μ is the *location parameter*; and $\Gamma(\cdot)$ is the Gamma function.

Estimate parameters in the Gamma function

If we know mean (denoted as u), variance (denoted as σ^2) and μ, then κ and θ can be obtained by

$$\theta = \frac{\sigma^2}{u - \mu}, \kappa = (\frac{u - \mu}{\sigma})^2 \tag{2}$$

Postulation

The mean and variance of travel times on a link depends on its free flow travel time τ^0 and the travel delay $\rho = \tau - \tau^0$; the location parameter μ depends only on τ_0.

Since ρ can be obtained from travel demand models, one can calibrate the above relationship using freeway data, then use the model to estimate mean and variance on arterial streets.

Linear regression

Linear regression model reads

$$
\begin{aligned}
u &= a_1\tau^0 + b_1\rho + c_1 && (3)\\
\sigma &= a_2\tau^0 + b_2\rho + c_2 && (4)\\
\mu &= a\tau^0 + b && (5)
\end{aligned}
$$

where $a, b, a_1, b_1, c_1, a_2, b_2$ and c_2 are coefficients to be estimated.

- ρ and τ^0 for all links (freeway and arterial) from a travel planning model prepared by Chicago Metropolitan Agency for Planning (CMAP).
- u, σ and μ are known on freeways and toll road, but unknown on arterial streets.

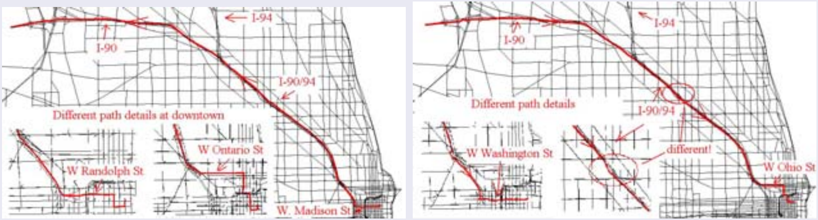
Linear regression results

time-of-day periods	Variance Model a_1	b_1	c_1	R^2
AM PEAK	0.309	0.870	0.580	0.444
PM PEAK	0.368	0.685	2.967	0.400
MIDDAY	0.283	1.076	2.040	0.346
OFF PEAK	0.178	0	-1.031	0.516

time-of-day periods	Mean Model a_2	b_2	c_2	R^2
AM PEAK	1.127	0.546	-2.056	0.910
PM PEAK	1.143	0.563	0.336	0.872
MIDDAY	1.100	0.630	-1.145	0.889
OFF PEAK	1.043	0.0000	-5.854	0.907

time-of-day periods	Location Model		
	a	b	R^2
AM PEAK	0.843	-4.106	0.958
PM PEAK	0.860	-3.533	0.964
MIDDAY	0.857	-3.608	0.956
OFF PEAK	0.831	-5.257	0.937

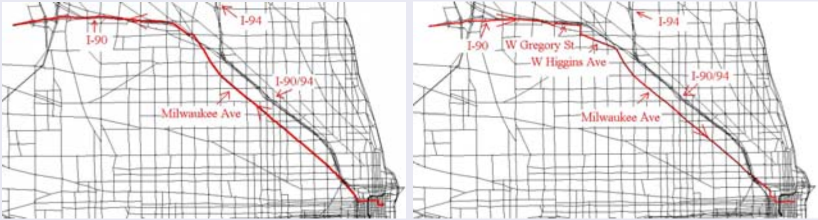
Downtown Chicago - the ORD Airport (Mid-of-Day)



(a) From downtown to ORD (b) From ORD to downtown

- For mid-of-day, FSD-admissible paths mostly use the freeway, as often suggested by Google Map or Yahoo maps.
- The differences among the paths are minor.

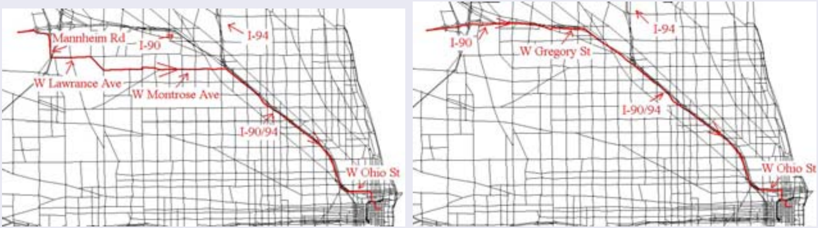
Downtown Chicago - the ORD Airport (Morning peak)



(c) from downtown to ORD (d) from ORD to downtown

- Drivers should stay away from the freeway if they wish to arrive on-time with high probability (95%).
- To arrive the airport with 95% probability, the reliable path requires a time budget of 33 minuets 57 seconds while using the freeway costs 37 minutes and 18 seconds to achieve the same reliability.

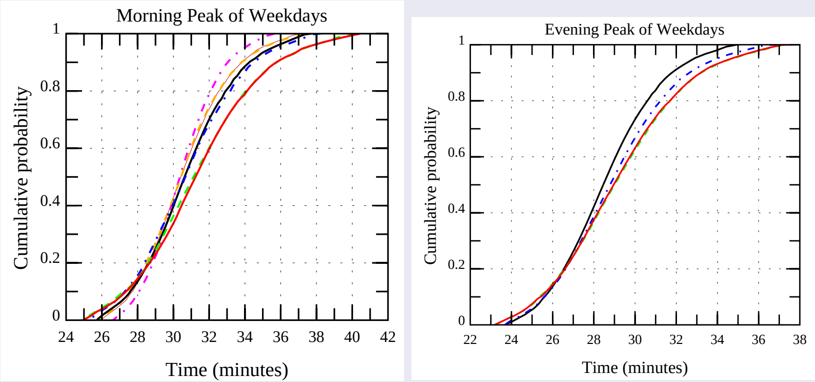
Downtown Chicago - the ORD Airport (Evening peak)



(e) 95% on-time arrival probability (f) 50% on-time arrival probability

- Motorists from the airport to the city should use arterial streets until they pass the merge of the two freeways.
- For 95% on-time arrival probability, the left path can save about 5 minutes comparing the right path.
- When 50% on-time arrival probability is required, the right path is slightly better (about 0.25 minutes).

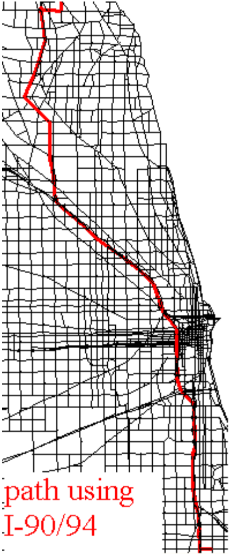
Distributions on FSD-admissible paths

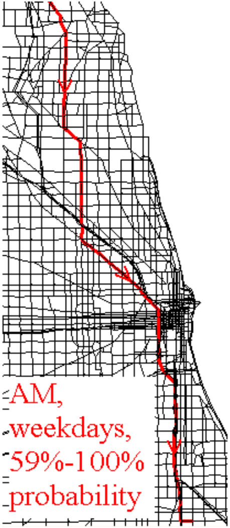


(g) Morning (h) Evening

Northshore - South suburbs (morning peak)

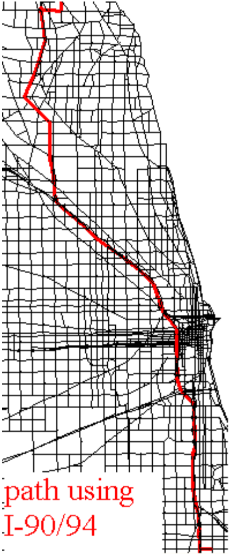
- For higher reliability motorists need to use various arterial streets until they are close to downtown Chicago, and then switch to the major freeway.

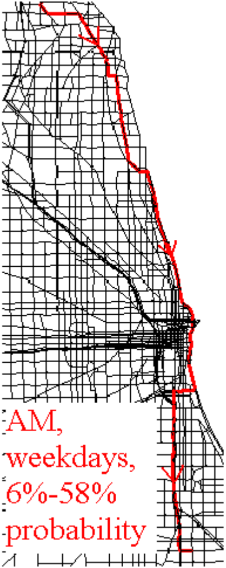




Northshore - South suburbs (morning peak)

- For lower reliability requirement, drivers can use another expressway known as Lake shore Dr.





Northshore - South suburbs (morning peak)

- For the mid-of-day and the evening peak periods, Lake Shore Dr. is more reliable.
- However, Lake shore Dr. is always preferred when traveling from South to North.

Computational performance

	Weekdays			Weekends		
	AM	Mid	PM	AM	Mid	PM
Downtown to ORD						
CPU time	29.58	18.69	16.58	12.25	19.14	8.50
# paths	7	5	4	1	5	1
ORD to downtown						
CPU time	29.58	23.70	14.58	15.69	15.36	28.02
# paths	6	2	2	1	2	4
Northshore to south suburbs						
CPU time	65.88	74.39	20.42	15.52	46.53	33.74
# paths	7	10	2	2	1	4
South suburbs to northshore						
CPU time	60.83	39.00	33.74	14.19	36.25	12.08
# paths	10	6	6	1	3	1

Computational performance (a sensitivity analysis)

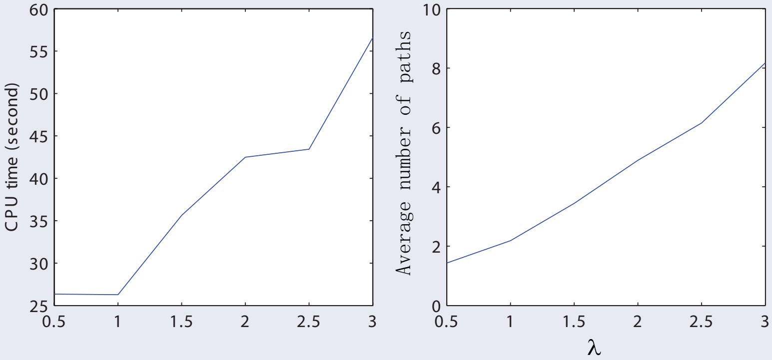


Figure: Impacts of variances on arterial streets on computational performance.

Summary

- General dynamic programming is used to formulate the reliable shortest path problem. Two theoretical results are essential:
  - Applicability of Bellman's Principle of Optimality
  - Acyclicity of admissible paths
- Reliable shortest path problem is NP-hard, but seems tractable when solved appropriately, even for very large problems
- Reliable route guidance does make a difference, and could generate substantial benefits in terms of time savings.
- Data availability remains a concern, particularly on arterial streets.

Possible extensions

- Consider higher-order stochastic dominance
  - Capture heterogenous risk-taking behavior
  - Reduce the number of non-dominant paths
  - Optimization atop of the non-dominant paths
- Application to traffic assignment and network design problems
- More efficient approximation algorithms
- Address more complete correlation structure
- Consider emerging data sources - such as GPS data, cell phone tracking, etc.

Acknowledgement

This research was funded by Commercialization of Innovative Transportation Technology (CCITT) from 2008 - 2009. The next stage of this research continues to receive funding from CCITT, and will also be jointly funded by National Science Foundation (NSF) and Illinois Department of Transportation (IDOT).

Resources

- A software tool, called Chicago Travel Reliability, or CTR, can be downloaded at http://translab.civil.northwestern.edu/nutrend/.
- We are currently conducting a survey to collect motorists' opinion about reliable routing. You could help us by providing your inputs (the survey can be accessed at the above URL).

Resources

Publication

1. Nie, Y., X.Wu, P. Nelson and J. Dillenburg (2009) Providing Reliable Route Guidance using Chicago Data, Technical Report #2009-001, CCITT.
2. X.Wu and Y. Nie (2009) Implementation issues in approximate algorithms for reliable a priori shortest path problem. Journal of the Transportation Research Board , 2091, 51- 60.
3. Nie, Y. and X. Wu. (2009) Reliable a priori shortest path problem with limited spatial and temporal dependencies. In the Proceedings of ISTTT-18, 169 - 196.
4. Nie, Y. and X. Wu. (2009) Shortest path problem considering on-time arrival probability. Transportation Research Part B, 43, 597-613.

Thank you!