University at Buffalo
School of Engineering and Applied Sciences

Incorporating social media data in search planning during Hurricane Irma evacuations

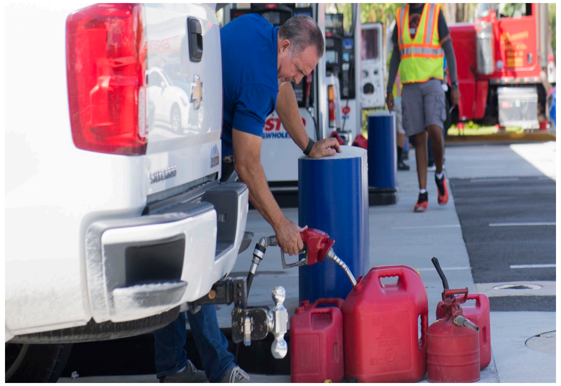
Abhinav Khare, Dr. Qing He, Dr. Rajan Batta
Department of Industrial and Systems Engineering, University at Buffalo

October 20th , 2019

Outline

- Motivation
- Objectives and Sub-problems
- Sub-problem 1 : Tweet Classification
- Sub-problem 2 : Event Localization
- Sub-problem 3 : Route Formulation
- Future Work

Motivation



- Gasoline is an extremely essential commodity pre/during natural disaster
  - Evacuations
  - Generators
- Surge in demand as people panic-by and hoard supplies → shortages
  - News outlets/ social media
  - Word of mouth
- Due to shortages, people are either stuck in high risks zones or take time in evacuations
- Such shortages became very prominent during the onset and post landfall of Hurricane Sandy (2011) and Hurricane Irma (2017).

Motivation

- According Florida Department of Transportation, during Irma demand of gasoline went up by 150%.

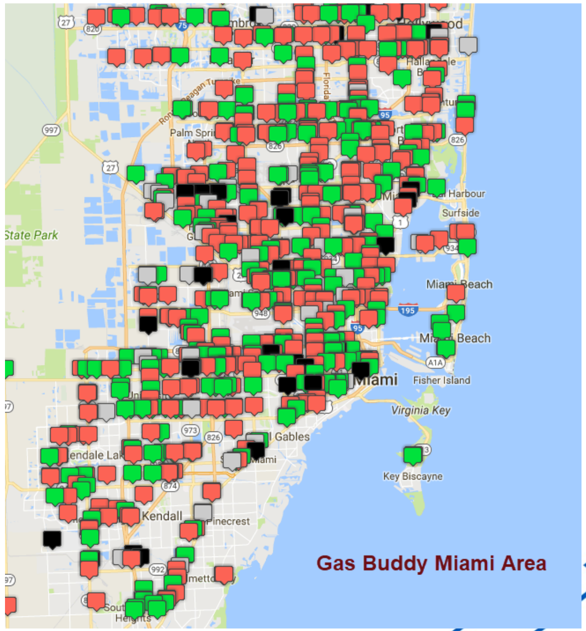
- There was enough gasoline at the ports to replenish the stations and satisfy the demands.

- There were not enough drivers and vehicles. They were brought in from Arizona later.

- People tweeted about these shortages like the following:

  1. “um so is there anywhere in town that still has gas”
  2. “i m wasting gas driving around trying to find it”
  3. “needed gas lines are crazy irmahurricane”
  4. “4 gas stations later and i finally got gas”

- These tweets contained information about shortage and geo-location of tweet has some association with locations of gas-station out of gas.



Gas Buddy Miami Area

Objective

- Question we try to answer:

Can we use the social media posts as inputs and output a strategy of gasoline search during evacuations (for individuals using social media)?

- We have successfully built a methodology for this problem.

- We believe this methodology can be generalized to other applications like searching for water, food and essentials in case of shortage .

- This method can potentially increase efficiency of evacuation and preparation during Hurricane onset.

Challenges and Sub-problems

Sub-problem 1: Need to identify tweets about shortage

- Twitter data is difficult to process and classify. It is unstructured, noisy and huge
- A tweet contains max 140 characters, informal, contain abbreviations and spelling mistakes.
- Classifying tweets for a specific item like identifying gasoline shortage has never been done.
- Identifying important features for this classification task is a novel and unique question

Sub-problem 2 : Need to geo-localize actual shortage to individual gas stations using locations of tweets

- Location and time of tweets not equivalent to the location and time of gas shortage or gas stations being talked about In the tweet.
- Spatial & temporal lag between the origin of the shortage & the tweet about shortage is an uncertain quantity.

Sub-problem 3 : Need to produce a search strategy given uncertainty about locations of shortage.

- Having inferred probabilistic information about gas stations without gasoline , we need to model and solve a search problem on a graph with a probability of finding the entity associated with each vertex.

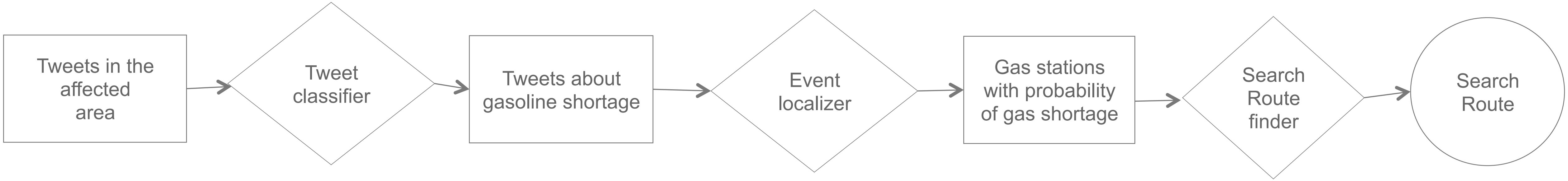

Tweets in the affected area
Tweet classifier
Tweets about gasoline shortage
Event localizer
Gas stations with probability of gas shortage
Search Route finder
Search Route

Figure: Flowchart of data processing pipeline for the gasoline search tool

Sub-problem 1 : Tweet Classification

Literature Review of previous Tweet Classifiers for Disaster Management

(Authors,year)	Labels	Classifiers	Features
(Verma et al, 2011)	(1) Situation awareness/ not, (2) Personal/impersonal, (3) Formal/ informal, (4) Subjective/objective	(1) Na ̈ive Bayes (2) Max-Entropy	(1) Unigrams and their raw frequency, (2) Bigrams and their raw frequency (3) Part-of-speech tags (4) Subjectivity of tweets (5) Register of tweets (6) Tone of tweet.
(Yang et al, 2013)	(1) Mitigation, (2) Preparedness, (3) Response, (4) Recovery	(1) Na ̈ive Bayes (2) Random Forest	(1) Unigrams
(Imran et al, 2013a), (Imran et al, 2013b)	(1) Personal, (2) Informative and (3) Other. (1) Caution and Advice, (2) Casualty and Damage, (3) Donations, (4) People, (5) Information Sources, (6) Others.	(1) Na ̈ive Bayes	(1) Unigrams, (2) Bigrams, (3) Part-of- speech tags, (4) Length of tweets, (5) Hasthags and (6) Emoticons
(Imran et al, 2014b)	User-defined categories of information like (1) Needs, (2) Damage etc.	(1) SVM	(1) Unigrams, (2) Bigrams, (3) Part-of- speech tags, (4) Length of tweets, (5) Hasthags, (6) Emoticons
(Ashktorab et al, 2014)	(1) Casualty/Infrastructure Damage or not	(1) K-nearest Neighbors, (2) Decision trees, (3) Naive Bayes, (4) Logistic Regression, (5) Latent Dirichlet allocation	(1) Unigram

Literature Review : Did not find classifier that could be directly used for our task.

(Authors,year)	Labels	Classifiers	Features
(Stowe et al, 2016a)	(1) Relevant, (2) Non-relevant, (1) Reporting, (2) Sentiment, (3) Information, (4) Action, (5) Preparation, (6) Movement,	(1) Classifiers of (Verma et al, 2011), (2) SVM	(1) Time of the tweet, (2) If the tweet was a retweet, (3) URL as a lexical feature, (4) Context of the tweet using unigrams/previous two tweets, (6) Word-embeedings, (7) Tag from the 4 classifiers, by (Verma et al, 2011).
(Ghosh et al, 2017)	(1) Resources available, (2) Resources required, (3) Medical Resources available, (4) Medical Resource required, (5) Requirements/availability of resources at specific locations, (6) Activities of various NGOs/ government organizations, (7) Infrastructure damage and restoration reported	(1) Naive Bayes, (2) SVM, (3) Decision Trees, (4) Random Forest (5) Gradient Boosting (6) Adaboost	(1) Unigrams with tf-idf scores, (2) Bi-grams with tf-idf scores, (3) K-nearest neighbour votes, (4) Length features, (5) List of units, (6) Availability related verbs (7) Medical related verbs, (8) Medical words, (9) Locational words.
(Stowe et al, 2018a)	(1) Relevant and (2) Non-relevant (1) Reporting (2) Sentiment (3) Information (4) Action (5) Preparation (6) Movement (1)Evacuation and (2) Not Evacuation	(1) Multi-Layer Perceptron, (2) Convolutional Neural Network	
(Alam et al, 2018a)	(1) Relevant and (2) Non-relevant	Graph Based Semi-supervised Learning with Convolution Neural Networks	

Data Description

- Our data one million tweets from Florida during the period 6-15 September 2017.
- 1048575 rows and 41 columns that include TWEET ID, TWEET TEXT, USER ID , DATE, HASHTAG, LATITUDE, LONGITUDE, BOUNDING BOX

Summary Statistic	Values
Number of Tweets Collected	1,048,575
Number of Unique Twitter Users	111,801
Period of Data Collection	6th Sept 2017- 15th Sept 2017
Date of Irma Landfall in Florida	9th Sept 2017
Number of tweets prior to Irma landfall in Florida	456,530
Number of tweets during Irma in Florida	151,792
Number of tweets post Irma in Florida	440,253
Number of Gas Related Tweets (labeld manually)	4070

Tweet Filtering

1. "Gasoline-related" tweets are the filtered out from the huge corpus tweets.
   - Used keyword search : any word that contains the "string" gas is a possible keyword
   - Used regular expression with the grep package in R for the keyword search.
   - Hashtag search : ^gas finds words starting with "gas" and also finds words that contain the string "gas"
   - Tweet search : \\bgas finds sentences with "gas" anywhere in the sentence
   - Results from tweet and hashtag search.

2. Tweet cleaning for removing noise and uniformity (for classification)
   - Removed user names, links, punctuations, tabs, general whitespaces, stopwords, and numbers.
   - Changed words to stem words and to lower case

Tweet Filtering

- Key words found using regular expressions:

gasoline, gas, gasinmiami, gaspricefixing, gasstation
gasservice, gastateparks, gasshortage, gasoil,
gastation, gaswaste, nogas, outofgas, findgas

- From 1 million to 4070 gasoline-related tweets were filtered out.

Tweet Classification

- Classified “gasoline-related” into “gasoline-shortage tweets” & “ non gasoline-shortage tweets”
- Classifier type: SVM
- Features type:
  1. Unigrams or words
  2. Sub-topics identified using Latent Topics identified using 4 kinds of topic models
     - i. Correlated Topic Models (CTM),
     - ii. LDA -1 : Latent Dirichlet Allocation (LDA) using Variational Expectiation-Minimization algorithm (VEM),
     - iii. LDA -2 : LDA using fixed VEM (VEM Fixed),
     - iv. LDA- 3 : LDA using Gibbs Sampling (Gibbs)
- In tweets about gasoline shortage , we saw there were few repeating themes. We wanted to explore if these subtopics can serve as identifier for the overarching topic of gas shortage.

Methodology – Tweet Classfication

Performance of SVM using topics and unigrams (varied word frequency threshold, number of topics = 5, train/test = 70/30

Term frequency threshold	Number of words	Precision	Recall	F score
5	937	0.9406566	0.7136015	0.8115468
6	797	0.9604142	0.7884615	0.8659845
7	710	0.9503121	0.7620137	0.8458096
10	519	0.9692899	0.7715618	0.8591967
20	282	0.9636684	0.7667436	0.8540008
50	109	0.9721385	0.7757009	0.862881
100	38	0.9722495	0.7793427	0.8651735
350	5	0.9852071	0.7894737	**0.8765465**

Methodology - Tweet Classification – Irma Results

- Performance of SVM using topics and unigrams (varied number of topics, word frequency threshold= 350, train/test = 70/30

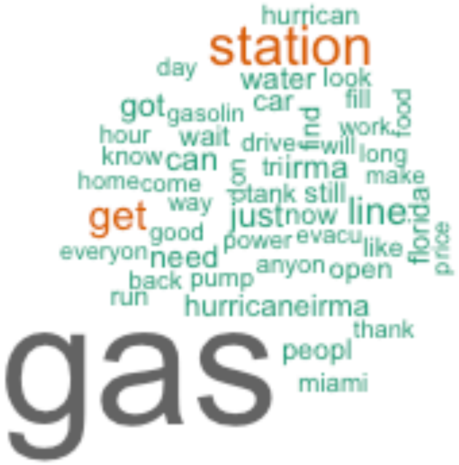
Number of topics	Precision	Recall	F score
2	0.9634378	0.7670813	0.8541196
4	0.9831247	0.7612366	0.8580682
5	0.9852071	0.7894737	**0.8765465**
6	0.9503142	0.7903614	0.8629887
10	0.9722495	0.7793427	0.8651735
10	0.9755556	0.71388	0.8244524
15	0.9733728	0.7878813	0.8708593

Methodology - Tweet Classification – Irma Results

- The best F1 score was obtained using 5 unigrams and 5 topics.
- The 5 best unigrams were gas, get, line, out, station.

5 topics identified using CTM topic modeling techniques

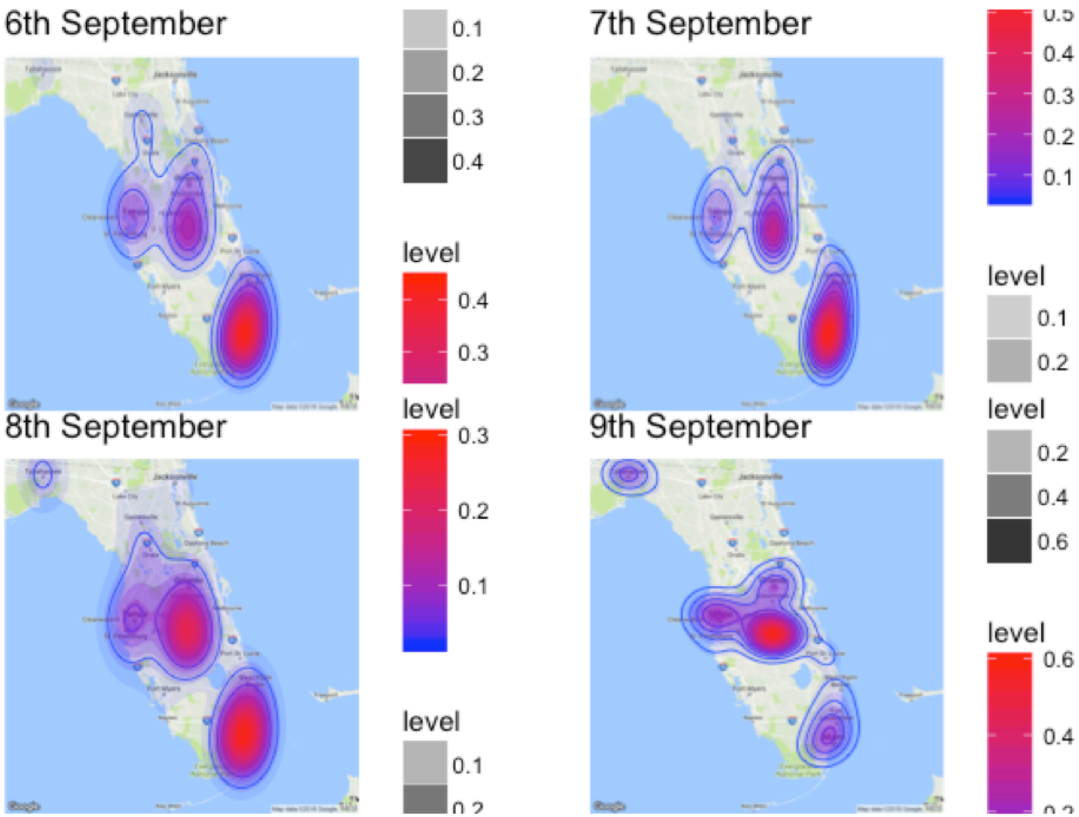
CTM				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
gas	station	gas	gas	gas
cannot	gas	no	station	price
find	need	station	wait	high
know	hurricaneirma	line	line	got
where	close	miami	irma	irma



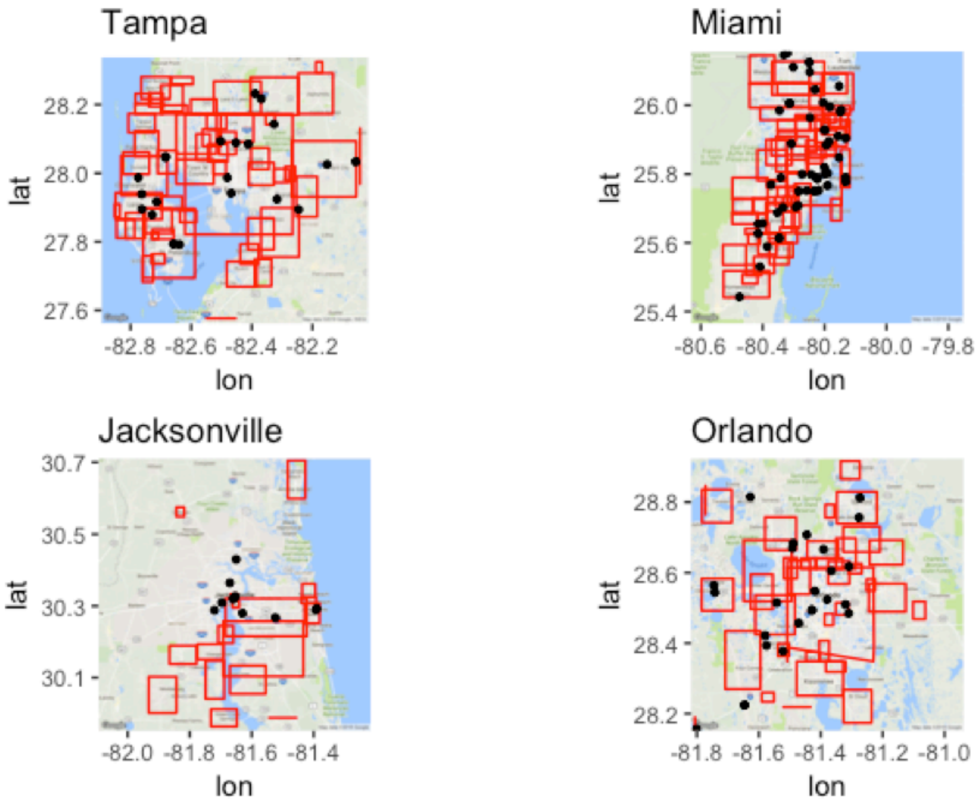
Word cloud of top 50 unigrams

Sub-problem 1 : Event Localization

Spatio-temporal Visualization of Tweets about Gasoline Shortage



Heat map of gasoline shortage tweets

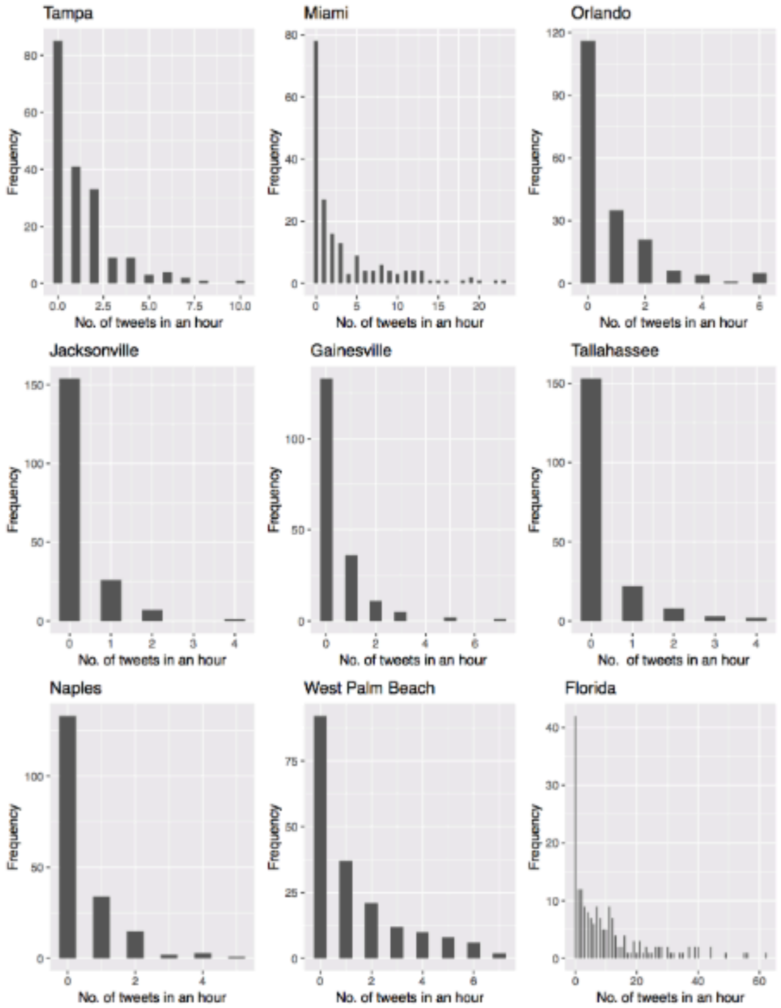


City-wise geo-location of gasoline shortage tweets

Hourly Arrival of gasoline-shortage tweets followed a Poisson Distribution

Results of Chi-square, VM, KS tests for Poisson Distribution

City	Lambda	Chi-sq p-value	VM p-value	KS test p-value
Tampa	1.281915	0.1894	0.000554025	0.0006129
Miami	3.356383	0.3073	0.000729358	2.22E–16
Orlando	0.7765957	0.1064	8.96E–05	0.09341
Tallahassee	0.2925532	0.3513	6.06E-09	0.5038
Jacksonville	0.2340426	0.4963	4.86E–09	1
Gainsville	0.4787234	0.1764	1.42E-06	0.6744
West Palm Beach	1.303191	0.2044	9.10E-04	0.01205
Naples	0.462766	0.2094	1.40E-06	0.5038
Florida	10.23936	0.1989	0.000347264	3.33E–15



Histogram for number of tweets in an hour

Literature Review of previous Event Localization Methods

(Author, year)	Methodology for Tweet/Event Localisation
(Sakaki et al, 2010)	Location of earthquake using Kalman filtering
(Middleton et al, 2013)	Location of event using Named Entity recognition in the tweet
(Singh et al, 2017)	Location of tweet predicted based on historical locations of the users and a Markov model
(Unankard et al, 2015)	Location of event using Named Entity recognition in the tweet
(Smith et al, 2017)	Location of flooded area using geocoded tweet combined with the simulations from a hydrodynamic model
(Kumar and Singh, 2019)	Location information of event extracted from tweet text using Convolutional Neural Network

Literature Review of previous Event Localization Methods

- Localisation method which used named-entity-recognition were not useful for our application as most of the tweets did not have location information in the textual content (Middleton et al, 2013; Unankard et al, 2015)
- (Sakaki et al, 2010)'s method to infer the temporal information of the earthquake event served as the motivation
- They modelled the time between time of the quake and the time of the tweet as an exponential random variable.
- We had found that arrival of tweets about shortage in a city followed a Poisson distribution. If arrival of gas shortage observation is assumed Poisson, then time between onservation and tweet follows an exponential distribution.
- Therefore, we modeled the time and dsatnce between observation of shortage at a gas station and a tweet about it as exponential random variable.

Bayesian Inference to Infer Probability of Shortage

- O_{it} stands for a Bernoulli random variable for observation of gasoline shortage at station i and time t
- $S_{it+\Delta t}$ stands for Bernoulli random variable for shortage at station i at time (t + Δt)
- L be the random variable for location of tweet and
- T be random variable for time of tweet, then following conditional probability distributions:
- CPD 1: $P_{O_{it}}(O_{it} = o_{it}) = \{0.5 \;\; if \;\; o_{it} = 1\,, 0.5 \;\; if \;\; o_{it} = 0\,\} \qquad \forall\, i \in V, t \in T$
- CPD 2: $P(tweet\ at\ location\ L | O_{1t} = 1\,, O_{it} = 0 \;\; \forall\, i \in V\backslash\{1\}) = \; \mu\Delta d e^{-\mu\Delta d} \;\; \forall\, \Delta d \; > 0$
- CPD 3: $P(tweet\ at\ time\ t + \;\Delta t \;| O_{1t} = 1\,, O_{it} = 0 \;\; \forall\, i \in V\backslash\{1\}) = \; \lambda\Delta t e^{-\,\lambda\Delta t} \;\; \forall\, \Delta t \; > 0$
- CPD 4: $P_{S_{it+\Delta t}}(S_{it+\Delta t}|O_{it}) = \{P_{O_{it}}(o_{it}) - \; \alpha\Delta t \;\; if \;\; o_{it} = 1\,, P_{O_{it}}(o_{it}) + \alpha\Delta t \;\; if \;\; o_{it} = 0\,\} \quad \forall\, i \in V, t \in T, \Delta t > 0$

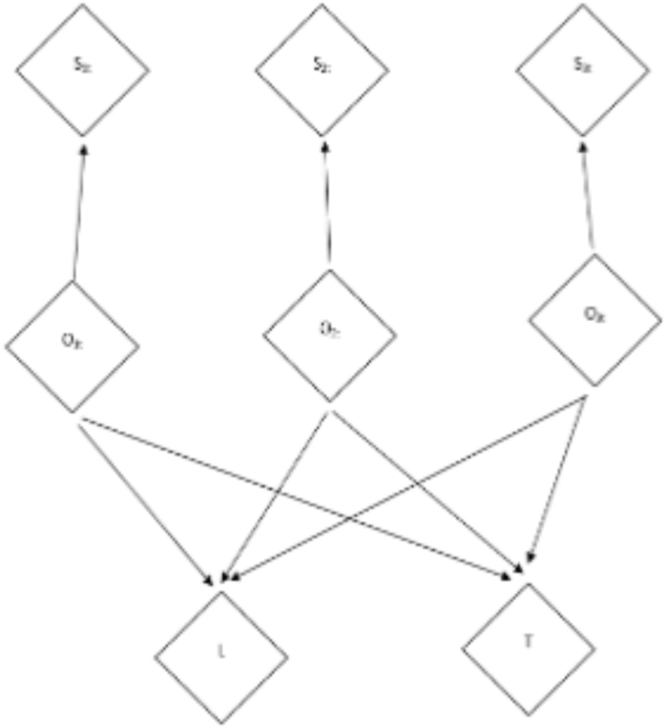


Fig. Snippet of the Bayesian network to infer gas shortage at eacl gas station

Event Localizer: Bayesian Inference to Infer Probability of Shortage at a Gas Station

- The CPDs were discretized and using variable elimination algorithm marginal probabilities of $S_{it+\Delta t}$ were calculated using the evidence from T and L giving us the probabilities of shortage at different gas stations at different times

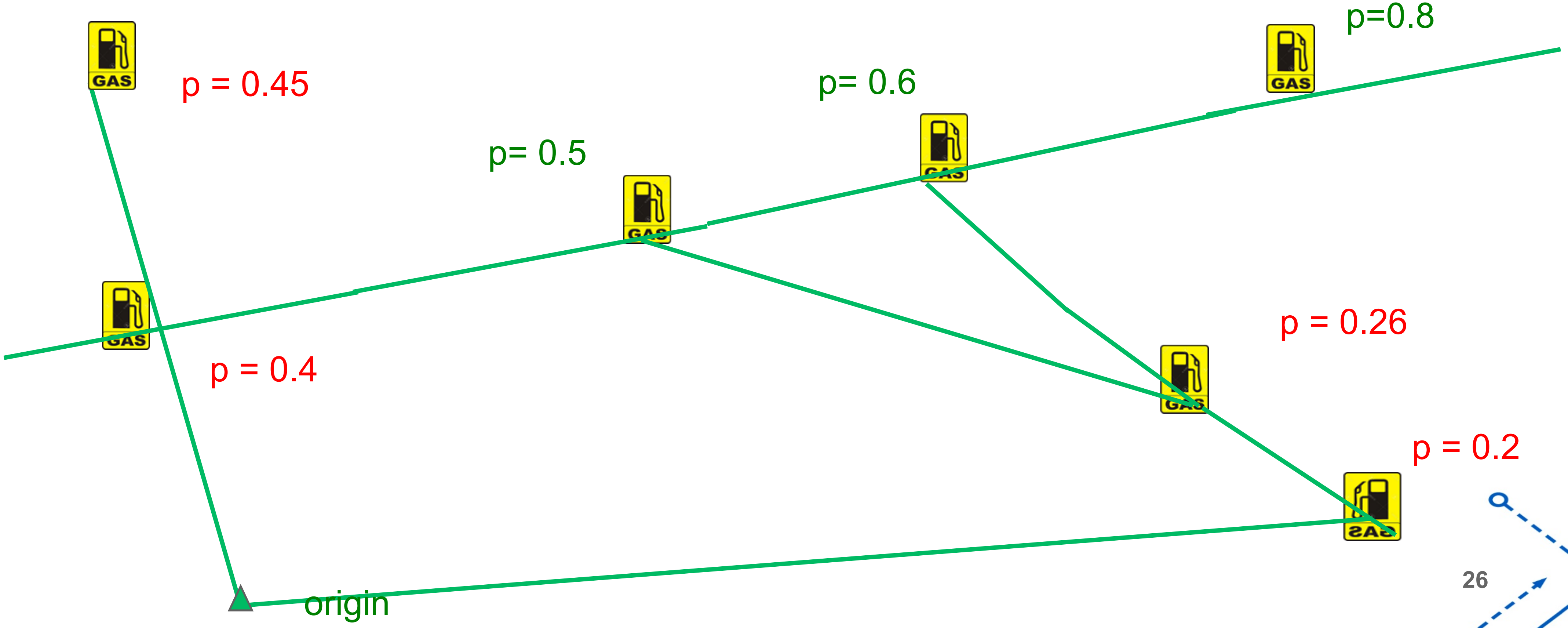
Table : Probability of gasoline shortage at different stations at different times

	190 SW 8th St, Miami	5701 Memorial Hwy, Tampa	4138 W Oak Ridge Rd, Orlando	10044 Atlantic Blvd, Jacksonville
Sept 8th, 9AM	0.53	0.5	0.57	0.51
Sept 8th, 11AM	0.55	0.55	0.61	0.51
Sept 8th, 1 PM	0.76	0.57	0.63	0.51
Sept 8th, 3 PM	0.72	0.58	0.62	0.53
Sept 8th, 5 PM	0.85	0.61	0.66	0.53
Sept 8th, 7 PM	0.85	0.67	0.67	0.53
Sept 8th, 9 PM	0.87	0.73	0.75	0.57

Sub-problem 3 : Search Problem

Search Problem

- The gasoline search problem can be explained as the problem of finding the path which minimizes the time for searching an entity on a graph with a finite probability of finding the entity at each vertex.



Search Problem

Parameters

- p_j , is the probability of finding gas at node j.
- d_{ij} is the distance between node i and node j.
- t_{ij} is the travel time between node i and node j.
- f is the amount of fuel left in the vehicle.
- m is the mileage of the vehicle.

Decision variables

- $x_{ijk} = 1$, if arc (i,j) is the k^{th} arc in the search path or else 0.
- $y_i = 1$, if node i lies in the search path or else 0.
- where,
- $i,j \in V = (0,1,,,,,n-1)$
- $k \in N = (1,,,,,n)$

Search Problem

- **Model 1 (Decide the order in to visit gas stations to minimize expected time for finding gas)**

Min

$$\sum_{k=1}^{n}(\prod_{m=1}^{k-1}\prod_{i=0}^{n-1}\prod_{j=0,i\neq j}^{n-1} p_j^{x_{ijk}}(1-p_j)^{x_{ijm}})(\sum_{l=1}^{k}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijl}t_{ij})$$

Expected search time given a path

St.

$$\sum_{k=1}^{n}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijl}d_{ij} = f * m$$

Length of path constrained by fuel left

$$\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijk} = 1 \ \forall\, k \in N$$

Only one arc can be the kth arc in the path

$$\sum_{j=1}^{n-1} x_{0j1} = 1$$

1st arc starts at the origin

$$\sum_{i=0}^{n-1} x_{ijk} = \sum_{i=0}^{n-1} x_{ji(k+1)} \ \forall\ k \in N, j \in V \setminus\{0\}$$

Destination of kth arc is the origin of k+1th arc

$$y_j = \sum_{k=1}^{n}\sum_{i=0}^{n-1} x_{ijk} \ \forall \ \ j \in V \setminus\{0\}$$

A vertex is the path if an arc enters it

Graph Search Problem

- **Linearize the non-linear objective function and use CPLEX or develop heuristic**

$$\mathit{Min}\ \sum_{k=1}^{n}\Big(\prod_{m=1}^{k-1}\prod_{i=0}^{n-1}\prod_{j=0,i\neq j}^{n-1} p_j^{x_{ijk}}(1-p_j)^{x_{ijm}}\Big)\Big(\sum_{l=1}^{k}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijl}t_{ij}\Big)$$

$$\sum_{k=1}^{n}\Big(\log\Big(\prod_{m=1}^{k-1}\prod_{i=0}^{n-1}\prod_{j=0,i\neq j}^{n-1} p_j^{x_{ijk}}(1-p_j)^{x_{ijm}}\Big)\Big(\sum_{l=1}^{k}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijl}t_{ij}\Big)$$

$$\sum_{k=1}^{n}\Big(\log\Big(\prod_{m=1}^{k-1}\prod_{i=0}^{n-1}\prod_{j=0,i\neq j}^{n-1} p_j^{x_{ijk}}(1-p_j)^{x_{ijm}}\Big) + log\Big(\sum_{l=1}^{k}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijl}t_{ij}\Big)\Big)$$

$$\sum_{k=1}^{n}\Big(\sum_{m=1}^{k-1}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{k-1}(x_{ijk}\log p_j + x_{ijml}\log(1-P_j)\Big) + \log\Big(\sum_{l=1}^{k}\sum_{i=0}^{n-1}\sum_{j=0,i\neq j}^{n-1} x_{ijl}t_{ij}\Big)\Big)$$

Search Problem

- **Dynamic Programming Formulation**

Recursive Bellman Equation

$$M_x = Min\,(\,1 - p_c)M_c + \; +p_c\,d_{xc} \;\; \forall\, c \in V \;\; - S$$

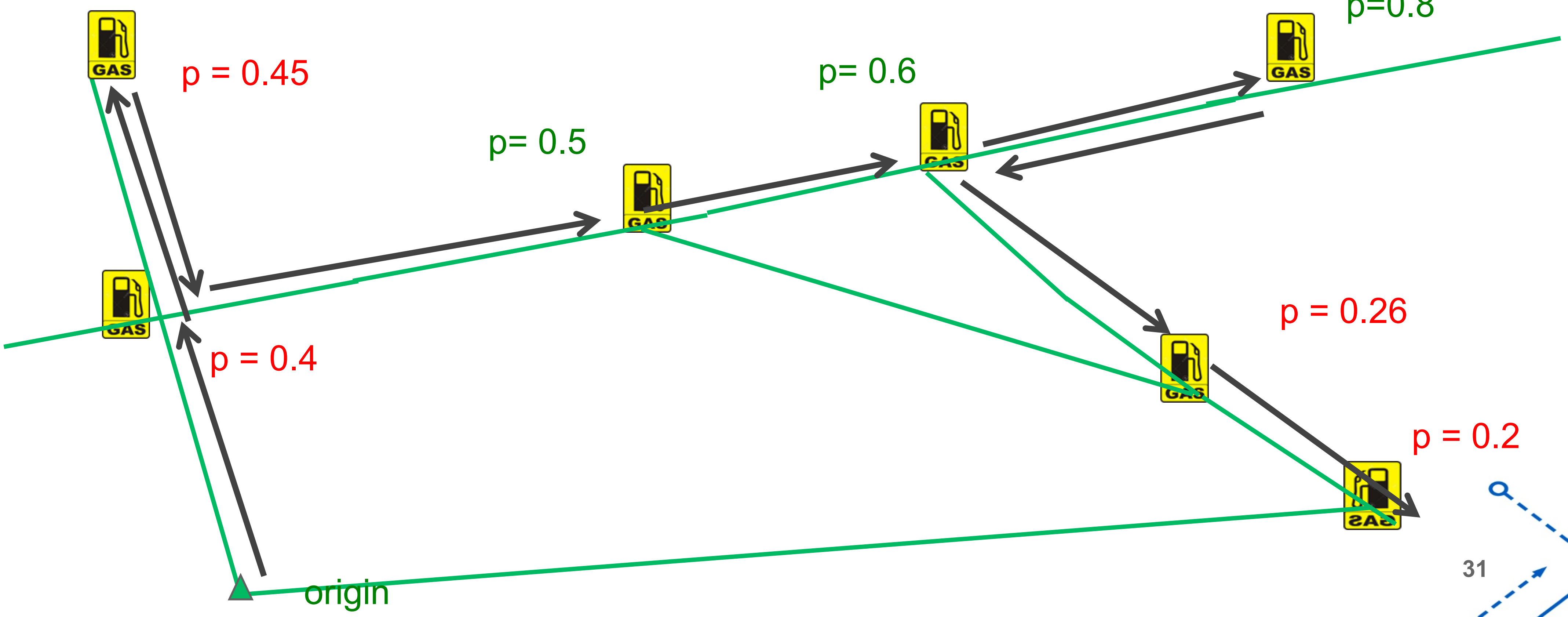
Where

M_x is the distance travelled from node x and to the end of the path

S is the set of nodes covered uptil node x

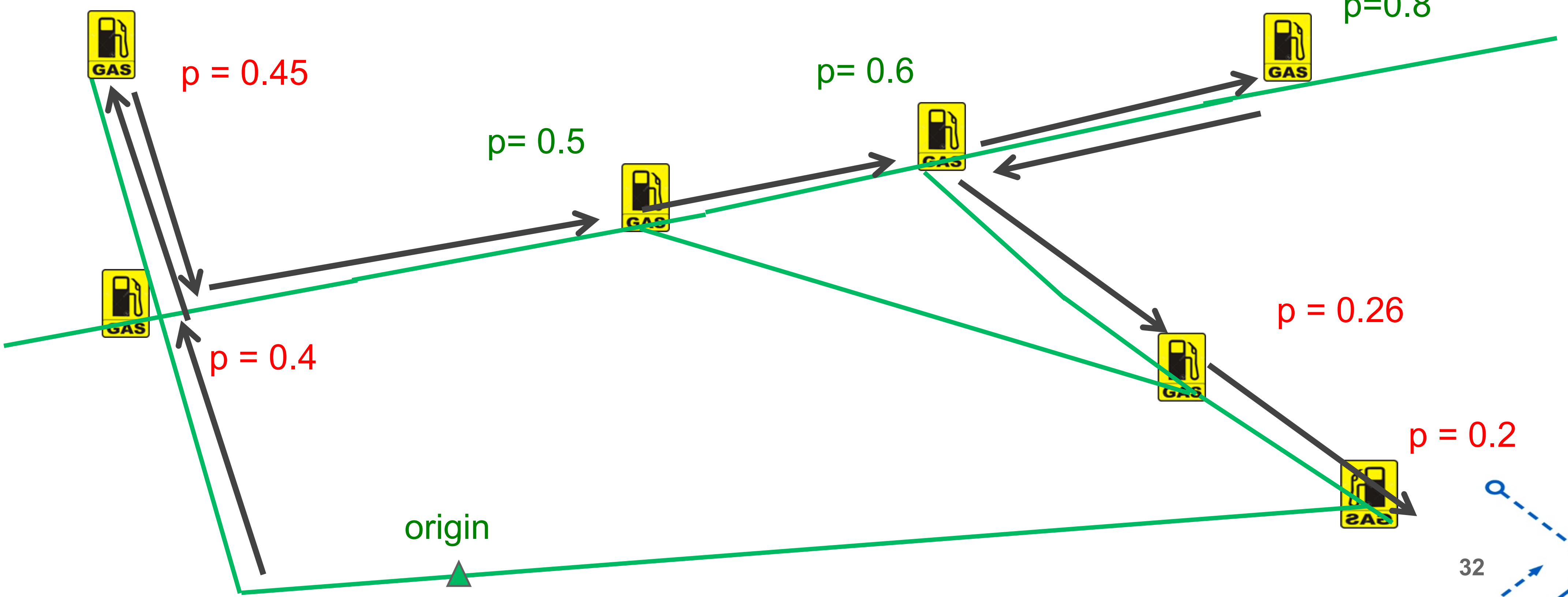
Results on Toy Problem

Original solution



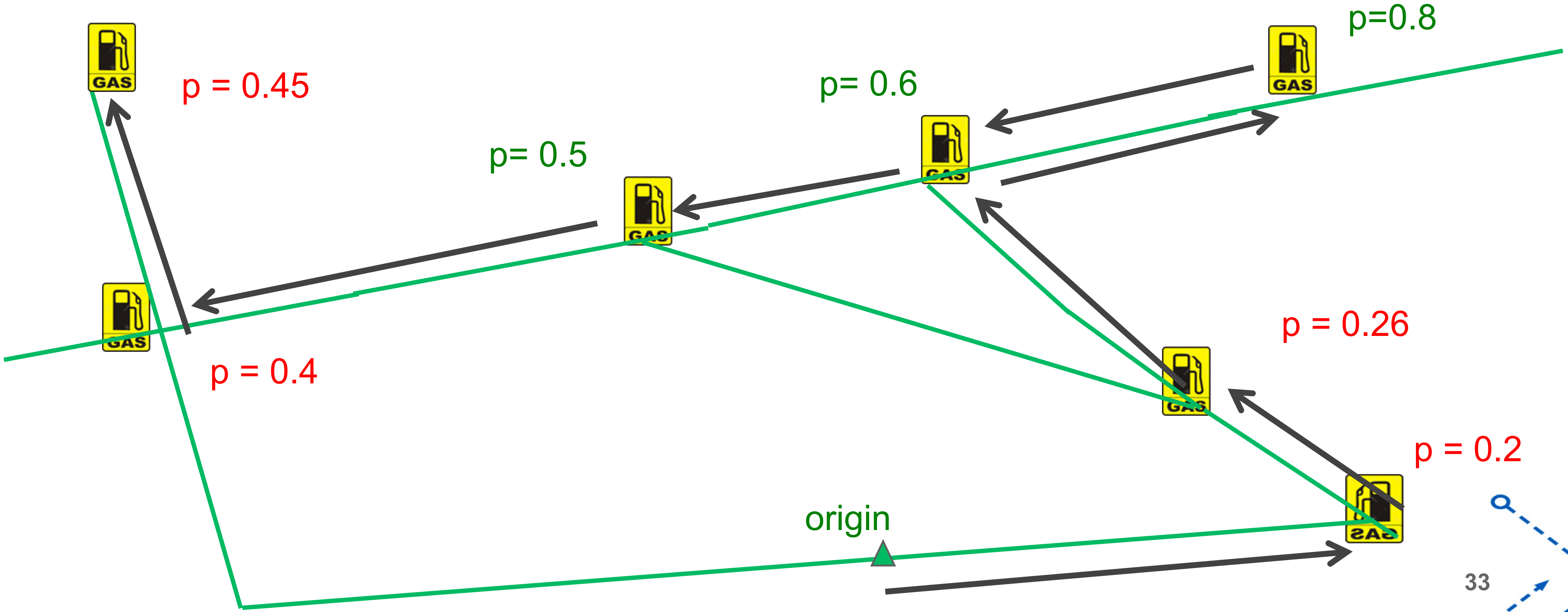
Results on Toy Problem

Changing origin



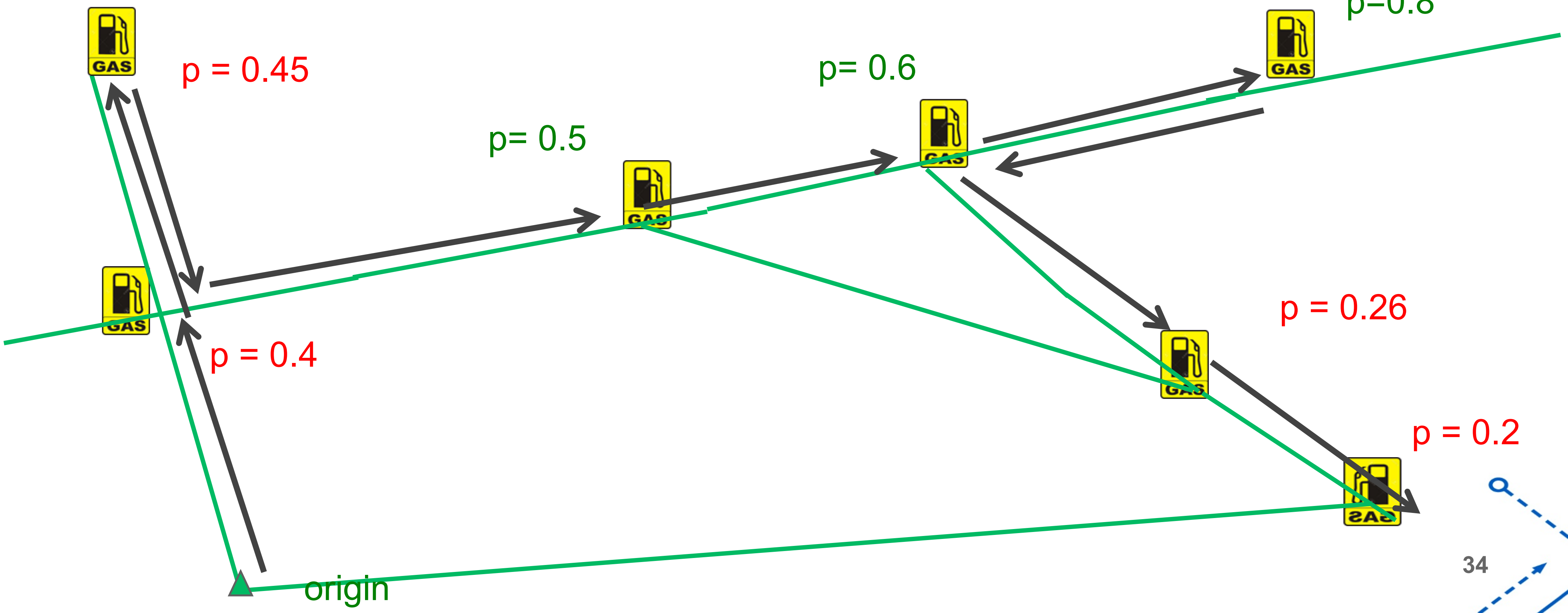
Results on Toy Problem

While changing origin there is a point on the arc at which the direction switches



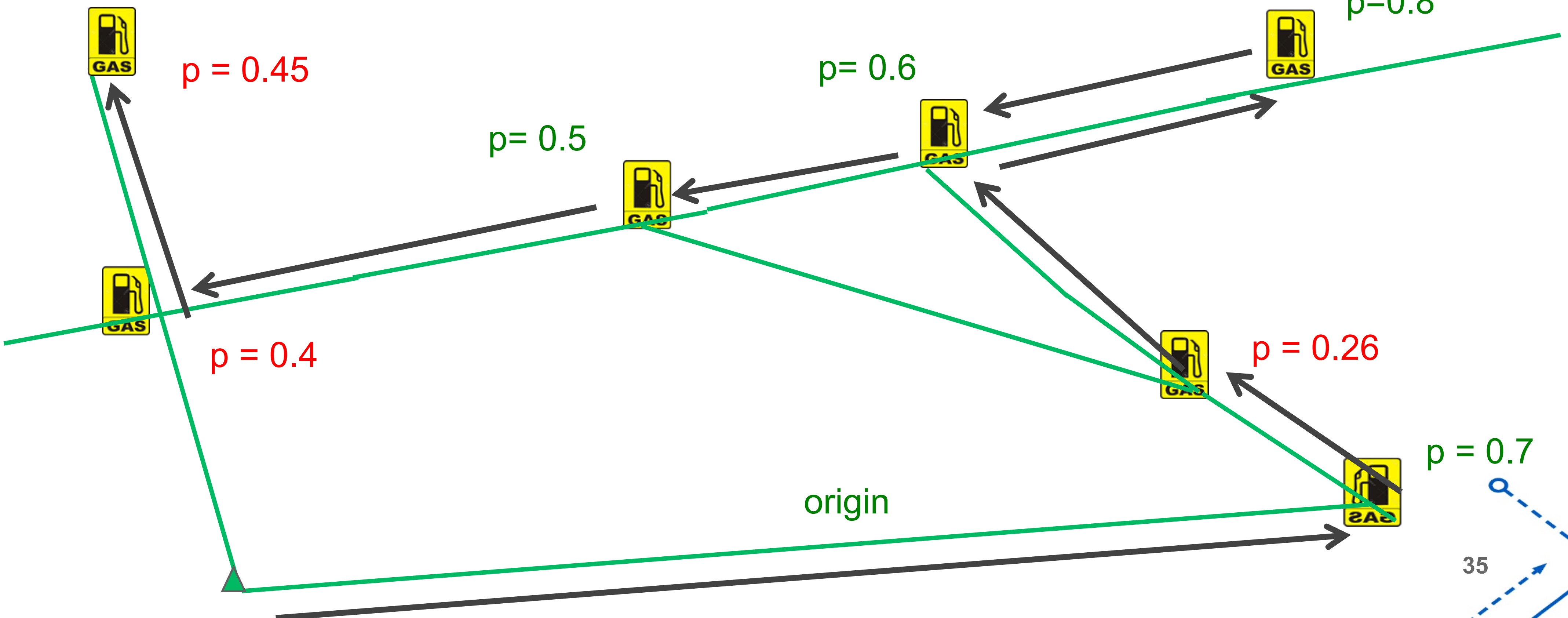
Results on Toy Problem

Original solution



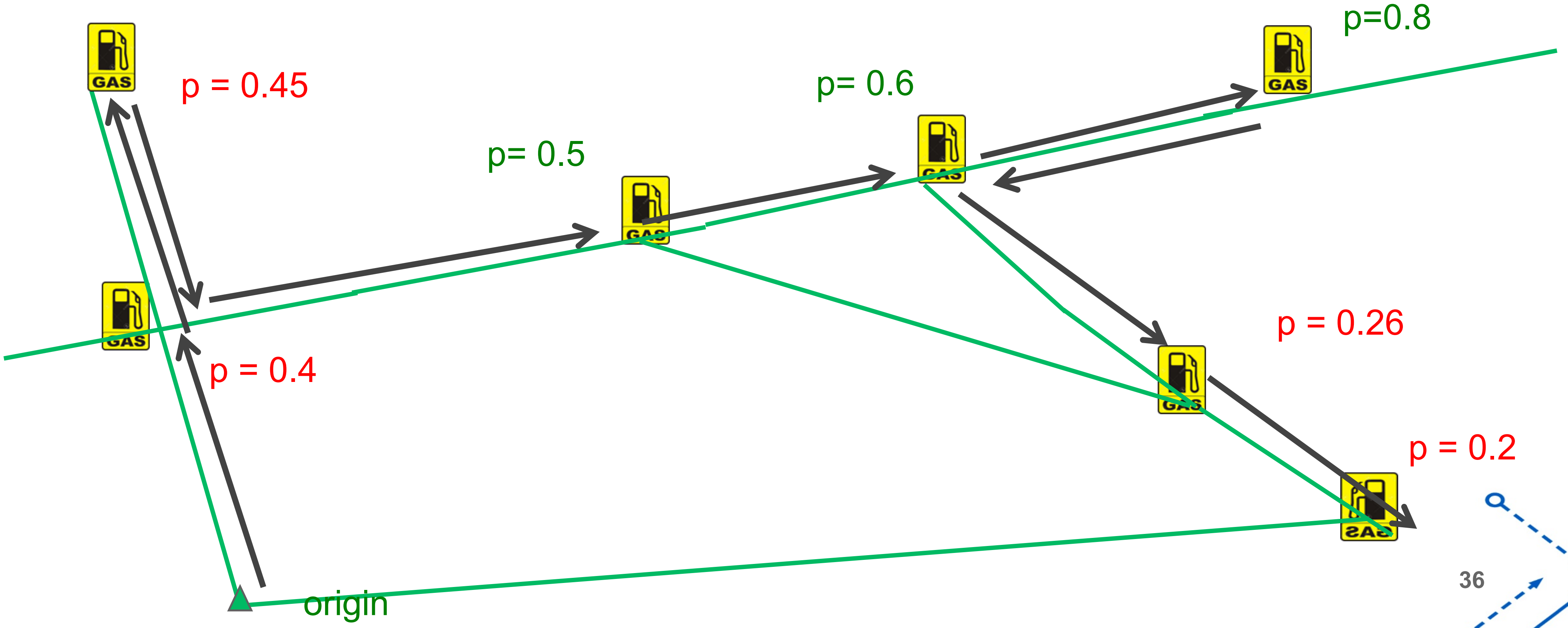
Results on Toy Problem

Path shifts if we change probabilities as well



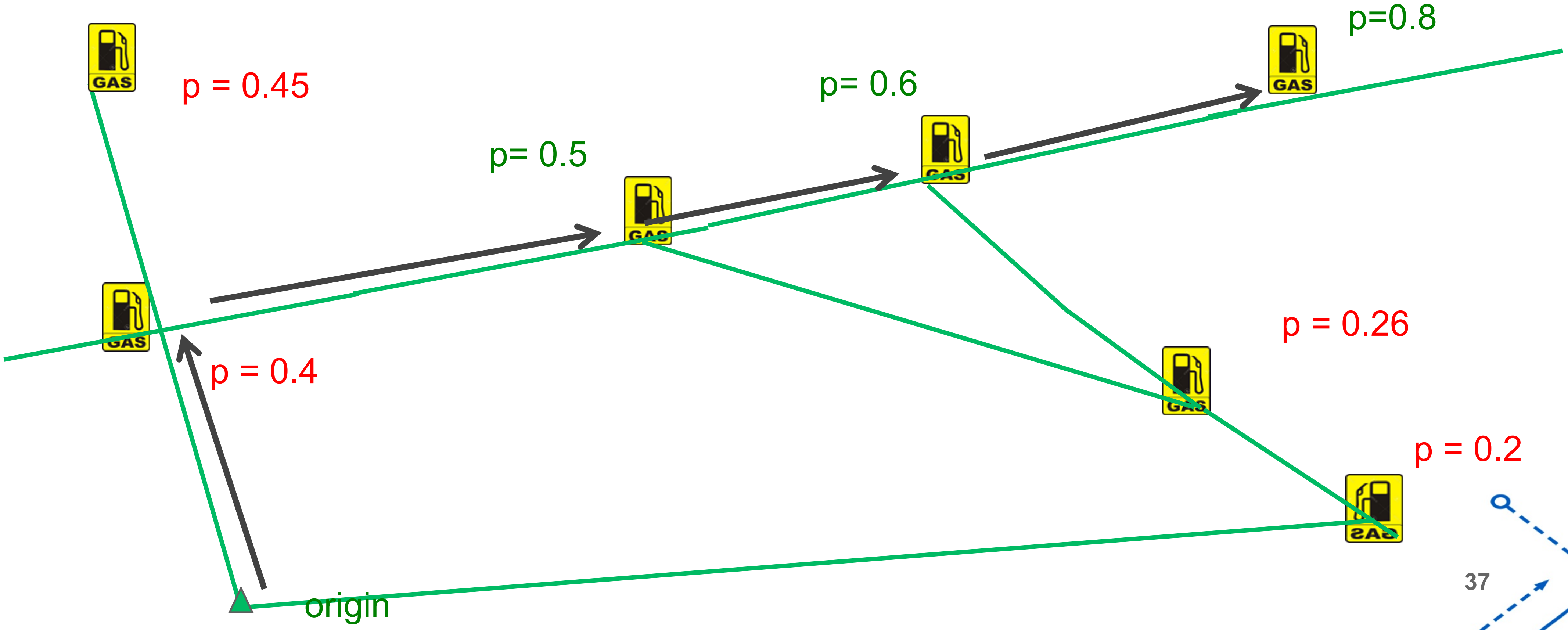
Results on Toy Problem

Original solution



Results on Toy Problem

The number of vertices reduces and order of visiting changes if amount of fuel is less



Discussion and Future work

- For real sized problems CPLEX and dynamic programming are inefficient.
- We are developing a heurestic to solve this problem efficiently.
- Current version of the problem has constant probabilties
- However, the probability at each vertex should update when the vehicle arrives at a vertex.
- Also, new tweets arrive as a vehicle is searching which can provide further updates in probabilities.
- A dynamic version of the problem can be solved with bayesian updates in probabilites and change in search paths

Thank you