Markov bases and subbases for bounded contingency tables

Ruriko Yoshida
Dept. of Statistics, University of Kentucky

Joint work with F. Rapallo

`polytopes.net`

		Serum Cholesterol (mg/100ml)						
	Blood	1	2	3	4	5	6	7
	Pressure	$<$ 200	200-209	210-219	220-244	245-259	260-284	$>$ 284
1	$<$ 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
2	117-126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
3	127-136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
4	137-146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
5	147-156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
6	157-166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
7	167-186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
8	$>$ 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source : [Cornfield, 1962]

Data on coronary heart disease incidence in Framingham, Massachusetts [Cornfield, 1962, Agresti, 1990]. A sample of male residents, aged 40 through 50, were classified on blood pressure and serum cholesterol concentration. $2/53$ in the (1,1) cell means that there are 53 cases, of whom 2 exhibited heart disease.

Imcomplete contingency table

Table 1: Effects of decision alternatives on the verdicts and social perceptions of simulated jurors.

	Condition						
Alternative	1	2	3	4	5	6	7
First degree	11	[0]	[0]	2	7	[0]	2
Second degree	[0]	20	[0]	22	[0]	11	15
Manslaughter	[0]	[0]	22	[0]	16	13	5
Not guilty	13	4	2	0	1	0	2

Source : [Vidmar, 1972]

This table refers to the possible effects on decision making of limiting the number of alternatives available to the number of a jury panel.

$[0]$ refers to the structural zero on the cell.

Independence model

Let $\mathbf{X} = \{X_{ij}\}$ be a $I \times J$ table $X_{ij} \in \mathbb{N}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

An observed table $X^{obs} = \{x_{ij}^{obs}\}$, $x_{ij}^{obs} \in \mathbb{N}$, and $1 \leq I, 1 \leq J$.

$$X_{ij} \sim Poi(\mu_{ij}) \text{ iid}$$

where $\mu_{ij} = \ln(\theta_{ij})$.

Consider the generalized linear model with a canonical linear predictor of the form:

$$\theta_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \lambda_{ij}^{RC}.$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

Independence model is a special case such that

$$\lambda_{ij}^{RC} = 0 \text{ for } 1 \leq i \leq I, 1 \leq j \leq J.$$

Hypothesis

The sufficient statistics for independence model include the row and column margins. Hence, the conditional distribution of the table counts given the margins is the same regardless of the values of the parameters in the model.

We have the following hypothesis test:

$$H_0 : \lambda_{ij}^{RC} = 0 \text{ no interaction.}$$
$$H_1 : \lambda_{ij}^{RC} \text{ not constant over all cells.}$$

Exact p-value computation

Let $\hat{\mathbf{X}}$ be the MLE of the data under the model. Then Pearson's χ^2 statistics is

$$f(X) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(\hat{X_{ij}} - X_{ij})^2}{\hat{X_{ij}}}.$$

An exact permutation test based on the χ^2 statistic is constructed as follows. The p-value of this test is:

$$p = E_{\mathbf{p}}[I_{\{f(\mathbf{X}) \geq f(\mathbf{x})\}}|\text{satisfying margins}]$$

where $\mathbf{x}$ is an observed table and $\mathbf{p}$ is the hypergeometric distribution.

In general we approximate the expected value by generating random draws from the hypergeometric distribution and estimate

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} I_{\{f(\mathbf{x}^i) \geq f(\mathbf{x})\}}$$

where N is the number of draws $\mathbf{x}^1, \cdots, \mathbf{x}^N$ iid from the hypergoemetric conditional on the sufficient statistics under H_0.

Note: This is the only possible method in situations where counts are very small or the number of tables satisfying margins is very small.

Question: How can we generate random draws from this distribution?

Answer: Apply Diaconis-Sturmfels algorithm to the MCMC technique. Diaconis-Sturmfels algorithm is the only method guaranteed to connect the MC.

What is a set of **moves** which connect all feasible contingency tables satisfying these margins?

For unbounded tables under independence model, we know the set of moves which connect all feasible contingency tables satisfying margins.

Note: We can generalize this problem by adding a bound for each cell of a table in addition to row and column sums.

Note: If some of the bounds are zeros, then it is a **incomplete** table, i.e., table with **structural zeros**.

Question 1: Finding a set of moves which connect all feasible bounded 2-way contingency tables satisfying the row sums and column sums.

Question 2: If we know these bounds are non-zero, i.e., it is not an incomplete table, then what is a set of moves connect all feasible bounded 2-way contingency tables satisfying the row sums and column sums?

Exact p-value computation

Note that the row sums and column sums are the sufficient statistics under H_0. For example, we have

				Total
	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	6
	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	6
Total	4	4	4	

and each cell is bounded by 2, i.e., $x_{i,j} \leq 2$ for $i = 1, 2$ and $j = 1, 2, 3$.

From the constraints we can set up the system of linear equations and inequalities.

e.g. For our 2×3 table, we have:

$$\begin{array}{cccccccc}
x_{1,1} & & & +x_{2,1} & & & = & 4 \\
& x_{1,2} & & & +x_{2,2} & & = & 4 \\
& & x_{1,3} & & & +x_{2,3} & = & 4 \\
x_{1,1} & +x_{1,2} & +x_{1,3} & & & & = & 6 \\
& & & x_{2,1} & +x_{2,2} & +x_{2,3} & = & 6 \\
& & & & & x_{i,j} & \in & \mathbb{Z}_+ \\
& & & & & x_{i,j} & \leq & 2
\end{array}$$

where $Z_+ = \{0, 1, 2, \cdots\}$.

By introducing slack variables we have the system of equations.

$$\begin{array}{ccccccccc}
x_{1,1} & & & +x_{2,1} & & & = & 4 \\
 & x_{1,2} & & & +x_{2,2} & & = & 4 \\
 & & x_{1,3} & & & +x_{2,3} & = & 4 \\
x_{1,1} & +x_{1,2} & +x_{1,3} & & & & = & 6 \\
 & & & x_{2,1} & +x_{2,2} & +x_{2,3} & = & 6 \\
 & & & & x_{i,j} & +y_{i,j} & = & 2 \\
 & & & & & x_{i,j} & \in & \mathbb{Z}_+
\end{array}$$

This is equivalent with $2 \times 3 \times 2$ tables with constraints $[A, C]$, $[B, C]$, $[A, B]$ for factors A, B, C, which would arise for example in case-control data with two factors A and B at three levels each.

In general, we can set up a system $\{x \in \mathbb{Z}_+^d | Ax = b\}$ for any tables.

Note: Thus, moves connect all integral points inside a feasible region $P_b = \{x \in \mathbb{R}^d | Ax = b, \, x \geq 0\} \neq \emptyset$.

What is a Markov Basis??

Suppose $P_b = \{x \in \mathbb{R}^d | Ax = b,\ x \geq 0\} \neq \emptyset$ and let M be a finite set such that $M \subset \{x \in \mathbb{Z}^d | Ax = 0\}$.

We define the graph G_b such that:

- Nodes of G_b are the lattice points inside P_b.
- We draw an undirected edge between a node u and a node v iff $u - v \in M$.

Definition : M is called a **Markov basis** if G_b is a connected graph for all b with $P_b \neq \emptyset$.

Why do we care?: A Markov basis is the only known set of moves which guarantees to connect all tables with any constraints.

Example

To make it simple we just removed bounds.

				Total
	? ? ?	? ? ?	? ? ?	6
	? ? ?	? ? ?	? ? ?	6
Total	4	4	4	

Table 2: 2×3 tables with 1-marginals.

There are 19 tables satisfying these margins. We counted using a software **LattE**.

$$\pm\ \begin{array}{|c|c|c|}\hline 1 & -1 & 0 \\ \hline -1 & 1 & 0 \\ \hline \end{array} \qquad \pm\ \begin{array}{|c|c|c|}\hline 0 & 1 & -1 \\ \hline 0 & -1 & 1 \\ \hline \end{array}$$

$$\pm\ \begin{array}{|c|c|c|}\hline 1 & 0 & -1 \\ \hline -1 & 0 & 1 \\ \hline \end{array}$$

There are 3 elements in a Markov basis modulo signs.

In fact such moves are called **basic moves**.

4	0	2
0	4	2

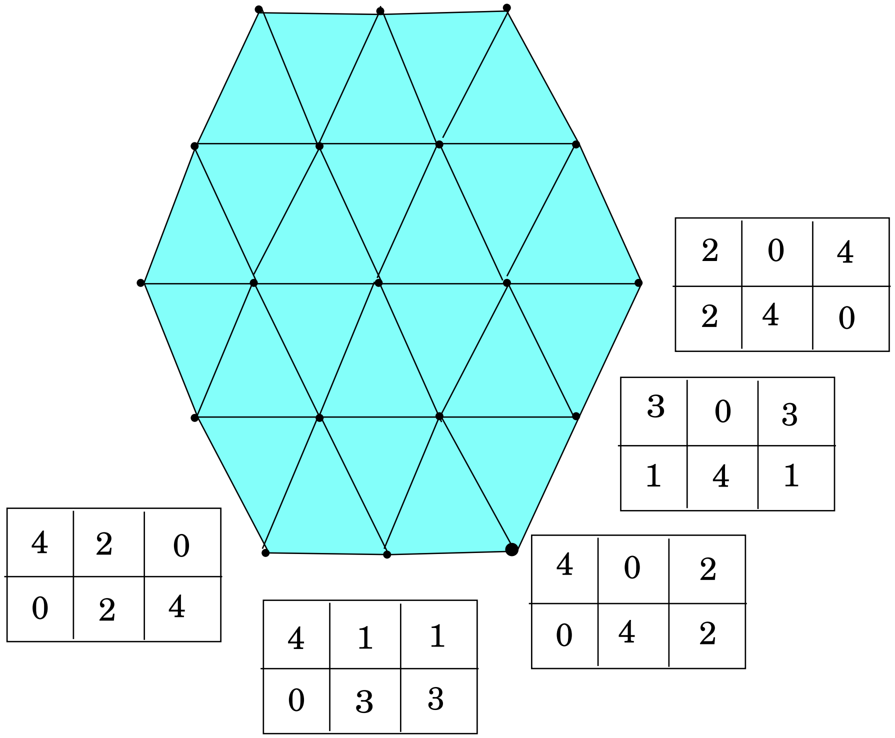
+

-1	0	1
1	0	-1

=

3	0	3
1	4	1

A table with the marginals plus an element of a Markov basis is also a table with the given marginals.



2	0	4
2	4	0

3	0	3
1	4	1

4	2	0
0	2	4

4	0	2
0	4	2

4	1	1
0	3	3

A Markov basis for 2×3 tables. An element of the Markov basis is a undirected edge between integral points in the polytope.

Fact: For any 2-way contingency tables with row and column sums (without bounds), we know that a set of basic moves forms a Markov basis.

However: If you add a constraint of bounds, then it is not necessarily true anymore.

For example,

0	0	1
1	0	0
0	1	0

0	1	0
0	0	1
1	0	0

with structural zeros in the diagonal cells are not connected by the basic moves.

Note: A Gröbner basis of a toric idea $\mathcal{I}_A$ associate to a matrix A with any term order is a Markov basis associate to a matrix A. So one can compute a Markov basis from a Gröbner basis of $\mathcal{I}_A$ with any term order.

Note: There are several nice software to compute Gröbner bases (such as **4ti2**).

However: Computing a Gröbner basis is very hard in general.

Notation

Without loss of generality, we represent a table by a vector of counts $\mathbf{n} = (n_1, \ldots, n_k)$. Let $\mathcal{X} = \{1, \ldots, k\}$ be the sample space of the contingency table. In the special case of two-way $I \times J$ tables, we will also denote the sample space with $\mathcal{X} = \{1, \ldots, I\} \times \{1, \ldots, J\}$.

The fiber of an observed table $\mathbf{n}_{\text{obs}}$ with respect to a function $T : \mathbb{N}^k \longrightarrow \mathbb{N}^s$ is the set

$$\mathcal{F}_T(\mathbf{n}_{\text{obs}}) = \left\{\mathbf{n} \mid \mathbf{n} \in \mathbb{N}^k \ , \ T(\mathbf{n}) = T(\mathbf{n}_{\text{obs}})\right\} \ .$$

When the dependence on the specific observed table is irrelevant, we will write simply $\mathcal{F}_T$ instead of $\mathcal{F}_T(\mathbf{n}_{\text{obs}})$.

In mathematical statistics framework, the function T is usually the minimal sufficient statistic of some statistical model.

Definition: A Universal Gröbner basis of an ideal is the Gröbner basis with respect to every term order.

Let a $s \times k$-matrix A_T be a configuration of T and $\mathcal{I}_{A_T}$ be a toric ideal associate with A_T.

Theorem [Rapallo and Rogantin, 2007] A Universal Gröbner basis of the toric ideal $\mathcal{I}_{A_T}$ is a Markov basis of bounded tables under the given model.

If we know a Universal Gröbner basis for A_T, then we can compute a MB for incomplete tables.

Computing a MB for incomplete tables

Let $\mathcal{X}_0 \subset \mathcal{X}$ be the set of structural zeros of the table, let T' be the function T restricted to $\mathcal{X}' = \mathcal{X} \setminus \mathcal{X}_0$ and let $\mathcal{I}'_{A_T}$ be the toric ideal associated with $A_{T'}$

Definition A Markov basis computed through a Universal Gröbner basis is a **Universal Markov basis**.

Theorem [Rapallo and Y., 2009] Let $\mathbf{n}$ be a contingency table and let $\mathcal{F}_T^{\mathbf{b}}$ be its bounded fiber under the bound $\mathbf{n} \leq \mathbf{b}$. Let $\mathcal{X}_0$ be the set of structural zeros. Then a Universal Markov basis for $\mathcal{F}_{T'}^{b}$ is obtained from a Universal Markov basis for $\mathcal{F}_T^{b}$ by removing the moves involving the cells in $\mathcal{X}_0$.

Example

Let us consider 4×4 contingency tables with fixed marginal totals. Without structural zeros, the Universal Markov basis is formed by 204 binomials: 36 moves involving 4 cells: 96 moves involving 6 cells: and 72 moves involving 8 cells.

Suppose that the cell $(1,1)$ is a structural zero. This kind of table is depicted below, where $[0]$ means a structural zero, while the symbol $\bullet$ denotes a non-zero cell.

$$\begin{pmatrix} [0] & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix}$$

Applying Theorem, we remove: 9 moves involving 4 cells: 36 moves involving 6 cells: and 36 moves involving 8 cells. The Universal Markov basis in this case has 123 moves.

Example cont...

Suppose now that the whole main diagonal contains structural zeros, as in the figure below.

$$\begin{pmatrix} [0] & \bullet & \bullet & \bullet \\ \bullet & [0] & \bullet & \bullet \\ \bullet & \bullet & [0] & \bullet \\ \bullet & \bullet & \bullet & [0] \end{pmatrix}$$

In this situation we remove: 30 moves involving 4 cells: 80 moves involving 6 cells: and 66 moves involving 8 cells. Finally, the Universal Markov basis has only 28 moves.

However, the Universal Gröbner basis of the toric ideal $\mathcal{I}_{A_T}$ is, in general, much bigger than a Gröbner basis of the toric ideal $\mathcal{I}_{A_T}$ with respect to a given term order. So in general it is very hard to compute.

Just to give the idea of such increase, we present in the following table the number of moves of the standard Markov basis for square $I \times I$ tables for the first I's.

	2	3	4	5	6	7
Standard Markov basis	1	9	36	100	225	441
Universal Gröbner basis	1	15	204	$3,940$	$113,865$	$4,027,161$

Thus, we consider the set of connecting moves.

Markov subbases

Definition: [Chen et. al., 2007] A Markov subbasis $M_{A_T,\mathbf{n}_{\text{obs}}}$ for $\mathbf{n}_{\text{obs}} \in \mathbb{N}^k$ and integer matrix A_T is a finite subset of $\ker(A_T) \cap \mathbb{Z}^k$ such that, for each pair of vectors $\mathbf{u}$, $\mathbf{v} \in \mathcal{F}_T$, there is a sequence of vectors $\mathbf{m}_i \in M_{A_T,\mathbf{n}_{\text{obs}}}, i = 1, \ldots, l$, such that

$$\mathbf{u} = \mathbf{v} + \sum_{i=1}^{l} \mathbf{m}_i,$$

$$0 \leq \mathbf{v} + \sum_{i=1}^{j} \mathbf{m}_i,\ j = 1, \ldots, l.$$

The connectivity through nonnegative lattice points only is required to hold for this specific $\mathbf{n}_{\text{obs}}$.

Note: $M_{A_T,\mathbf{n}_{\text{obs}}}$ for every $\mathbf{n}_{\text{obs}} \in \mathbb{N}^k$ and for a given A_T is a Markov basis $\mathcal{M}$ for A_T.

Markov subbases for tables with positive bounds

We first study Markov subbases $M_{A_T,\mathbf{n}_{\text{obs}}}$ for any bounded two-way contingency tables $\mathbf{n}_{\text{obs}} \in \mathbb{N}^k$ with positive bounds, i.e., no structural zeros, under independence model.

Theorem [Rapallo and Y., 2009] Consider $I \times J$ tables with row and column sums fixed and with all cells bounded. If these bounds are positive, then a Markov subbasis for the tables is the standard Markov basis for $I \times J$ tables with row and column sums fixed without bounds, i.e., the set of basic moves of all 2×2 minors.

Example

Consider now 4×4 tables with fixed row and column sums, and with all bounded cells.

The constraint matrix that fixes row and column sums in a 4×4 table gives a toric ideal with a $\binom{4}{2} \times \binom{4}{2}$ element Gröbner basis, i.e., a Markov basis is formed by the basic moves of the form $\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$ for all 2×2 minors of the table.

The full Markov basis for bounded tables has 204 moves. However, by the theorem above the Markov subbasis for this table is the standard Markov basis for a 4×4 table with fixed row and column sums fixed without bounds.

Markov subbases for incomplete tables

Now we study Markov subbases $M_{A_T,\mathbf{n}_{\text{obs}}}$ for any incomplete $I \times J$ contingency tables $\mathbf{n}_{\text{obs}} \in \mathbb{N}^k$ with positive margins, i.e., $A_T(\mathbf{n}_{\text{obs}}) > 0$, under independence model.

Without loss of generality, we can assume that all margins are positive because cell counts in rows and/or columns with zero marginals are necessary zeros and such rows and/or columns can be ignored in the conditional analysis.

Let $\mathcal{X} = \{(i,j) \mid 1 \leq i \leq I, 1 \leq j \leq J\}$ and let S be a non-trivial subset of $\mathcal{X}$.

Proposition [Aoki and Takemura, 2005] Suppose we have $I \times J$ tables with fixed row and column sums. A set of basic moves is a Markov subbasis for $I \times J$ contingency tables, $I,\ J \geq 4$, with structural zeros in only diagonal elements, i.e., (i.e., cells with indices in $S = \{(i, j) : i = j$ for $i = 1, \ldots, \min(I, J)\})$ under the assumption of positive marginals.

How about if cells in S are structural zeros, where S does not contain diagonals?

We consider 3×3 and 4×4 tables under independence model with all cells bounded. We assume row and column sums are positive. We have studied in which S all cells can be structural zeros in order for the standard Markov basis, i.e., the moves of the form $\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$ for all 2×2 minors of the table, to connect these bounded tables with positive conditions.

To analyze these cases we recall some definitions from commutative algebra:

An ideal $\mathcal{I} \subset \mathbb{R}[\mathbf{x}]$ is *radical* if

$$\{f \in \mathbb{R}[\mathbf{x}] \mid f^n \in \mathcal{I} \ \text{ for some } n\} = \mathcal{I}\,;$$

Let $\mathcal{I},\ \mathcal{J} \subset \mathbb{R}[\mathbf{x}]$ be ideals. The quotient ideal $(\mathcal{I} : \mathcal{J})$ is defined by:

$$(\mathcal{I} : \mathcal{J}) = \{f \in \mathbb{R}[\mathbf{x}] \mid f \cdot \mathcal{J} \subset \mathcal{I}\}\ ;$$

Let $\mathcal{I},\ \mathcal{J} \subset \mathbb{R}[\mathbf{x}]$ be ideals. The saturation of $\mathcal{I}$ with respect to $\mathcal{J}$ is the ideal defined by:

$$(\mathcal{I} : \mathcal{J}^{\infty}) = \{f \in \mathbb{R}[\mathbf{x}] \mid g^m \cdot f \in \mathcal{I},\ g \in \mathcal{J},\ \text{for some } m > 0\}\,;$$

Let $Z = \{z_1, \ldots, z_s\} \subset \mathbb{R}^k$. A lattice L generated by Z is defined:

$$L = \mathbb{Z}Z.$$

$M \subset \mathbb{R}^k$ is called a lattice basis of L if each element in L can be written as a linear integer combination of elements in M.

Theorem [Chen, Dinwoodie, and Y., 2008] Suppose $\mathcal{I}_M$ is a radical ideal, and suppose M is a lattice basis. Let $p = x_1 \cdots x_k$. For each index ℓ with $(A_T)_\ell > 0$, let $\mathcal{I}_\ell = \langle x_h \rangle_{(A_T)_{\ell,h} > 0}$ be the monomial ideal generated by indeterminates for cells that contribute to margin ℓ. Let $\mathcal{L}$ be the collection of indices ℓ with $(A_T \mathbf{n})_\ell > 0$. Define

$$\mathcal{I}_{\mathcal{L}} = \left(\mathcal{I}_M : \prod_{\ell \in \mathcal{L}} \mathcal{I}_\ell \right).$$

If

$$(\mathcal{I}_{\mathcal{L}} : (\mathcal{I}_{\mathcal{L}} : p)) = \langle 1 \rangle \qquad (1)$$

then the moves in M connect all the tables in $\mathcal{F}_T$.

Using this theorem we study incomplete 3×3 and 4×4 tables.

Incomplete 3×3 tables

If $|S| = 1$ or $|S| = 2$ then Equation in (1) holds. Thus, the 9 moves of the form $\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$ for all 2×2 minors of the table connect bounded tables. For $|S| = 3$, if $S = \{(1,1), (2,2), (3,3)\}$ after an appropriate interchange of rows and columns, i.e. there are 6 patterns of S, then Equation in (1) does not hold. Otherwise for other patterns of S, Equation in (1) holds. Thus, 9 moves connect bounded tables. For $|S| > 3$, if S contains $\{(1,1), (2,2), (3,3)\}$ after appropriate interchange of rows and columns, then Equation in (1) does not hold. Otherwise for other patterns of S, Equation in (1) holds. Thus, these 9 moves connect bounded tables.

Big open problem for Markov subbases for incomplete tables

We also consider 4×4 tables under independence model with all cells bounded. We assume row and column sums are positive. After an appropriate interchange of rows and columns, if we have structural zero constraints on all diagonal cells (i.e., cells with indices in $S = \{(i, j) : i = j$ for $i = 1, \ldots, I\}$), then Equation in (1) does not hold.

Using the previous proposition and examples we have the following problem.

Problem [Rapallo and Y., 2009] Suppose we have $I \times J$ tables with fixed row and column sums. What is the necessary and sufficient condition on S so that the set of basic moves connects all tables under the assumption of positive marginals?

Advertisement...

The Sepcial Session on Advances in Algebraic Statistics

Organized by Sonja Petrović and RY

2010 AMS Spring Southeastern Sectional Meeting

Lexington, KY, March 27–28, 2010 (Saturday – Sunday)

`http://www.ams.org/amsmtgs/2162_program_ss2.html#title`

Thank you....

The paper is available at http://arxiv.org/abs/0905.4841.